5-2015

GlyR3 regulation in *Clostridium thermocellum*

Jinlyung Choi
University of Tennessee - Knoxville, jchoi12@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Jinlyung Choi entitled "GlyR3 regulation in *Clostridium thermocellum*." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Chemical Engineering.

Chris D. Cox, Major Professor

We have read this dissertation and recommend its acceptance:

Eric T. Boder, Paul D. Frymier, Cong T. Trinh, Qiang He

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

GlyR3 regulation in *Clostridium thermocellum*

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Jinlyung Choi
May 2015

ACKNOWLEDGEMENTS

I would like to thank Chris D. Cox for his support and guidance during my Ph. D. program at the University of Tennessee. I would also like to thank Eric T. Boder, Paul D. Frymier, Cong T. Trinh and Qiang He for serving on my committee. Special thanks go to Gary S. Sayler for his support, knowledge and encouragement during this process. I would also like to thank a number of people at the Center for Environmental Biotechnology for their assistance in my research including: James Fleming, Alice Layton, Steven Ripp, Dan Williams, Scott Moser, FuMin Menn. I also offer my thanks to Logan Nester for his assistance in the lab.

I would also like to thank the U.S. Department of Energy funded Bioenergy Science Center and the Institute for a Secure and Sustainable Environment at the University of Tennessee for providing funding. I would also like to thank the Joint Genome Institute for analysis of RNA samples used in this project.

Special thanks also go to the family especially my parents, my wife (Lanying Ma) and my daughter (Arielle) for their support and existence.

ABSTRACT

Bio-ethanol from cellulosic biomass is a promising candidate as a liquid transportation fuel because of its high-energy content and the abundance of cellulose. Consolidated bioprocessing (CBP) helps to reduce the traditional 2-step process of bio-ethanol production into single-step to improve cost efficiency. A bacterium, *Clostridium thermocellum*, has a multi-enzyme complex for hydrolyzing cellulase, called the Cellulosome that enables the organism to have high rates of cellulose utilization. However, the ethanol yield of *C. thermocellum* needs to be improved in order to make consolidated bioprocessing with *C. thermocellum* commercially viable. It is essential to understand the regulation of carbohydrate-degrading enzyme activity to apply metabolic engineering, synthetic biology, and molecular biology techniques to strain improvement. GlyR3 is a protein that regulates the activity of carbohydrate active proteins in *C. thermocellum*. Recent studies have described how GlyR3 regulates the *celC* operon, which includes two carbohydrate-active enzymes (Newcomb, Chen, & Wu, 2007a). In this dissertation I investigate additional regulatory targets of the GlyR3 protein, which is a LacI family protein. First, it will be shown that GlyR3 regulates a gene downstream of *celC* operon, *manB*, explaining that GlyR3 not only induces its own expression under presence of laminaribiose, but in the case of *manB*, repress expression. Bioinformatics tools are then used to find putative GlyR3 binding sites in the whole genome in *C. thermocellum*. Electrophoretic mobility shift assay (EMSA) can show direct binding between GlyR3 and DNA

sequences of interest. Second, we are going to show that GlyR3 regulates genes that are located far from the *celC* operon. Reverse transcript (RT) –PCR and mRNA sequencing will be used to show *in vivo* changes in expression of genes in the whole genome resulting from changing levels of GlyR3 expression resulting from the addition of laminaribiose. This research reveals two distinct binding motifs for GlyR3, a celC-type and a manB-type, which are used to identify additional operons potentially regulated by GlyR3 and demonstrates a global regulatory role for GlyR3 in *C. thermocellum*.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Attachments

Additional File 1: rSEQ data validation

A PDF file containing the figure of rSEQ data validation

File name: Additional File 1.pdf

Additional File 2: Primers and probe

A PDF file containing the table of primers and probe

File name: Additional File 2.pdf

Additional File 3: EMSA

A PDF file containing the figure of EMSA

File name: Additional File 3.pdf

Additional File 4: Full list of gene expression change

An excel file containing the full list of gene expression change

File name: Additional File 4 Full list of gene expression change.xlsx

Additional File 5: New candidates

An excel file containing the full list of the new candidates

File name: Additional File 5 New candidates.xlsx

Additional File 6: Operon structure of candidates

A PDF file containing operon structure of candidates

File name: Additional File 6 Operon structure of candidates.pdf

Additional File 7: Gene expression profile of Cthe_0391 and Cthe_0392

A PDF file containing gene expression profile of Cthe_0391 and Cthe_0392

File name: Additional File 7.pdf

Additional File 8: Bacterial strain

A PDF file containing the table of bacterial strain

File name: Additional File 8.pdf

Additional File 9: Negative control for *in vitro* transcription assay

A PDF file containing the figures from the negative control experiment

File name: Additional File 9.pdf

Additional File 10: Operons and genes potentially regulated by GlyR3 identified using the celC and manB binding motifs and differential gene expression data

An excel file containing the full list of the newly found genes

File name: Additional File 10.xlsx

INTRODUCTION

Bio-Ethanol is a promising candidate of replaces petroleum-based fuels in the transportation sector.

The world's economic activities are dependent on the availability of economical energy sources. Energy sources such as petroleum and coal are a major cause of the acceleration of global warming since they produce carbon dioxide when they are burned. However energy demand is still increasing, especially in developing counties. So, it is important to develop renewable energy to replace fossil-based energy sources. As part of this effort, the U.S. government has set goals to increase the portion of the renewable energy among transportation fuels. (Sannigrahi, Ragauskas, & Tuskan, 2010; Solomon, Barnes, & Halvorsen, 2007)

There are many candidate technologies to replace fossil-fuel based transportation, such as hydrogen-fuel cells, electric vehicles, hybrid-electric vehicles, bio-ethanol and bio-diesel. However, vehicles powered by hydrogen-fuel cells seem difficult to commercialize in the near future. Electric vehicles are available, but their driving range is limited because of battery capacity and extended time of recharging. So, bio-ethanol could be a good candidate to use as a liquid transportation fuel. It is relatively safer than hydrogen to store and deliver. Bio-ethanol can be used in standard gasoline engines up to a concentration of 10% without modification and in concentrations up to 85% in specially designed engines. It doesn't required extended periods to fill the tank. It is advantageous on long-distance traveling with frequent refilling. Bio-ethanol produces carbon dioxide when it burns but the energy crops that are used for

bioethanol production fix carbon dioxide when they grow. So, it does not increase concentration of carbon dioxide in the air. (Solomon et al., 2007)

In addition to bio-ethanol, bio-butanol is considered as a second-generation biofuel. Researchers think bio-butanol has potential to replace gasoline in an internal combustion engine since it is more similar to gasoline than ethanol. Also, butanol can be produced from biomass. (Atsumi et al., 2010; Bhandiwad et al., 2013)

Most of bio-ethanol is currently produced from sugar and starch based materials such as corn and sugarcane (Mussatto et al., 2010). In 2008, more than 95 percent of bio-ethanol in the US was produced from corn. But, using corn-based ethanol as a transportation fuel can result in competition between food and fuel. In 2008 corn prices became too high due to the demand from bio-ethanol. Corn production can vary every year because of market demands and environmental conditions such as drought or flood. These factors could increase criticism that farmlands are used for fuel instead of food. Also, corn ethanol has low net energy balance because it cannot use the stover and stem portions of the plant, but only uses starch from the grain. Growing sufficient corn to satisfy a significant fraction of US transportation fuel demands would require huge quantities of land and water. Growing corn also contributes to soil erosion and loss of biodiversity. Corn also uses nitrogenous fertilizer and pesticides, which can create additional environmental problems. (Giampietro, Ulgiati, & Pimentel, 1997; Sannigrahi et al., 2010; Solomon et al., 2007)

Switchgrass is an example of a fuel crop that can be grown on marginally productive lands such as the side of a highway and may have less need for irrigation. Switchgrass, other grasses and trees are characterized by high cellulose content, which represents a vast potential energy resource. For these reasons, cellulosic ethanol has potential to be more sustainable (Lynd, Wyman, & Gerngross, 1999). Cellulosic ethanol can exploit underutilized resources such as idle land and waste biomass such as wood chips from furniture manufacturing. Also, it is possible to use almost all parts of the plant because cellulose is the major component of the plant cell wall. For example, cellulosic technology can make use of the corn stover, including the stem, leaf, and cob portions of the plant, compared to conventional technology that can only utilize the starchy grain portions of the plant. Switchgrass is another promising candidate as a bioenergy crop because it has high productivity per acre and grows in most parts of North America and can be grown on marginal croplands without fertilizer(David & Ragauskas, 2010). Cellulose is a major component of paper. Using waste paper for ethanol production can save demands on landfills. Researchers believe bio-ethanol from cellulose can solve the problem because it is the most abundant component in the plant biomass(Lynd, Weimer, van Zyl, & Pretorius, 2002b). The U.S. government wants to increase the renewable energy portion of the country's energy portfolio.

The major challenge of utilizing cellulose contained in plant cell walls is that it forms a stable structure that is difficult to breakdown to convert to ethanol

(Chundawat, Beckham, Himmel, & Dale, 2011). The cell wall is composed of cellulose, hemicellulose and lignin. The cellulose and hemicellulose components contain the sugars that can be fermented into biofuels; however, their stable structure in the cell wall greatly reduces their bioavailability. But lignin is a complex polymer of aromatic alcohols. It gives the cell wall strength but can inhibit enzyme activity to produce ethanol (Sannigrahi et al., 2010). (Figure 1)

C. thermocellum is a model organism to produce ethanol from cellulosic biomass

A conventional cellulosic bio-ethanol process includes two major steps. First, the sugars contained within cellulose and hemicellulose must be freed from the plant cell wall and broken down into short chains or sugar monomers. Second, a microorganism, such as yeast, ferments the sugars into ethanol or other biofuels. The first step in particular is time and energy consuming. A potential process improvement is to produce ethanol in a single processing step through consolidated bioprocessing using a bacterium that has many enzymes able to hydrolyze cellulose and ferment the resulting sugars into ethanol. *Clostridium thermocellum* is considered to be a model organism for consolidated bioprocessing. It is a thermophilic, anaerobic bacterium that has a cellulosome. The cellulosome is a multi-enzyme complex consisting of 1) cohesion, 2) dockerin, 3) scaffold and 4) catalytic subunits. The dockerin subunits connect the catalytic enzymes to cohesion subunits on the scaffoldin (Figure 2-1).

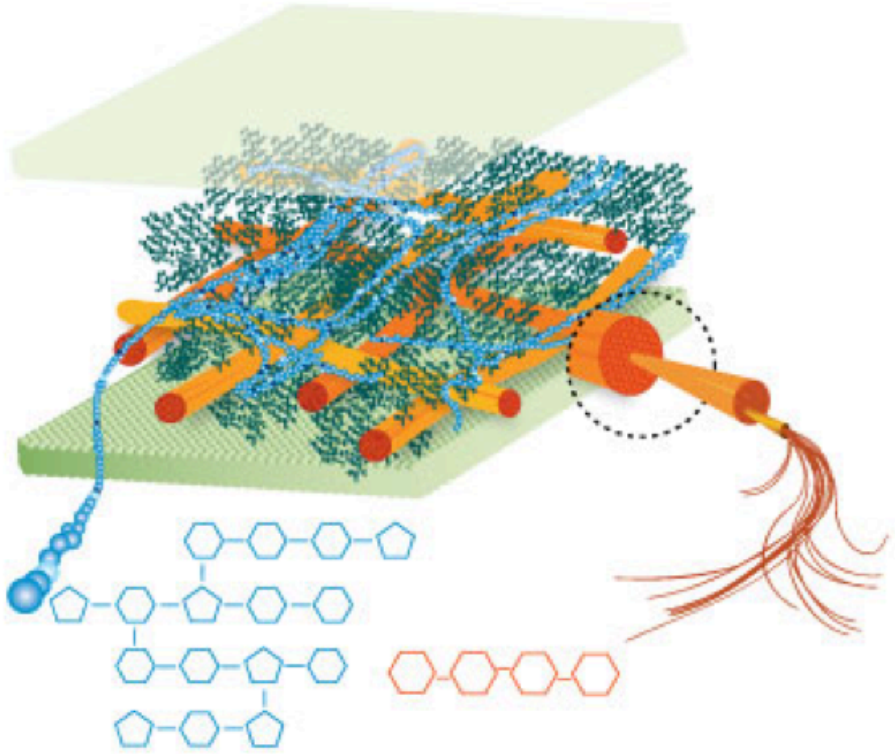


Figure 1 Cell wall structure

Green: cell wall, Orange: cellulose, Blue: hemicellulose, Dark green: lignin

(Ceres, 2014)

The anchoring scaffoldin is attached to cell and connected to the primary scaffoldin by an adaptor (Figure 2-2,3). The presence of specific catalytic units within the cellulosome depends on available carbohydrate substrates and other environmental factors, but the mechanism of regulation is not well researched yet. (Lynd, Zyl, McBride, & Laser, 2005)

Many researchers have studied this bacterium. However, relatively few studies have focused on gene regulation in *Clostridium thermocellum*. This study will focus on developing a better understanding of gene regulation in *C. thermocellum.* Such an understanding will contribute to improved strain engineering with the goal of increasing the efficiency of consolidated bioprocessing to realize a commercially viable process.

Example of gene regulation of Cellulosome related genes

Anti sigma factor regulation in *C. thermocellum*: Anti-sigma factors are trans-membrane proteins characterized by an extracellular carbohydrate binding module and a sigma-factor binding module extending into the cytoplasm. The anti-sigma factor binds the sigma factor when a specific polysaccharide is not present. Each anti-sigma factor has a binding affinity specific to a particular sigma factor. When the sigma factor is bound to the anti-sigma factor, the genes under its control are not expressed. However, the anti-sigma factor releases the sigma factor when a specific polysaccharide binds to the carbohydrate-binding module outside of the cell.

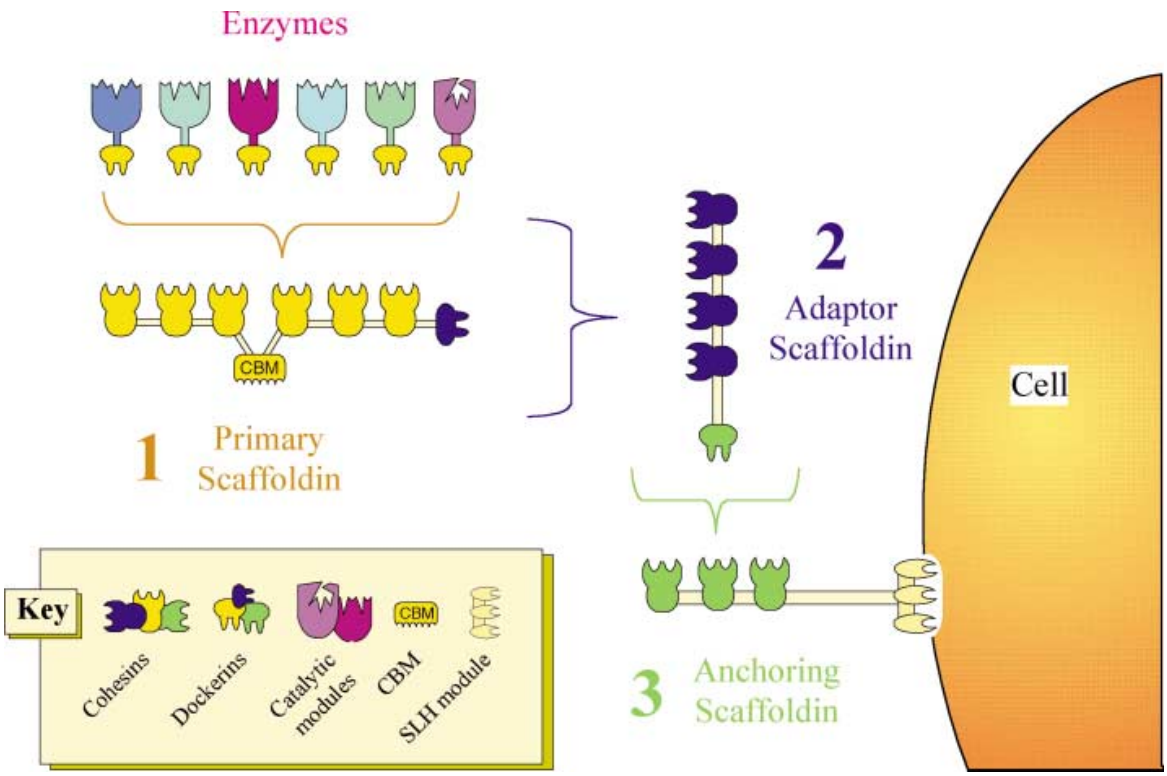


Figure 2 Structure of the cellulosome

(Bayer, Belaich, Shoham, & Lamed, 2004b)

Then the sigma factor can enable expression of the genes under its control.

Kahel-raifer et al. and Bahari et al. tested anti-sigma factors 1,3,5 6, and 24 of *C. thermocellum,* respectively. They incubated insoluble polysaccharides with anti-sigma factors, separated them by centrifugation, and determined the extent of binding by SDS-PAGE. They found anti-sigma factor 1 binds cellulose, 3 binds pectin, 5 binds arabinoxylan, 6 binds xylan and cellulose, and 24 binds cellulose. (Bahari et al., 2011b; Kahel-Raifer et al., 2010)

Nataf et al. found that anti-sigma factor (RsgI), which is located down stream of its sigma factor, binds to sigma factor I. Also the sigma factor and anti-sigma factor are regulated together. Isothermal titration calorimetry was used to confirm that the sigma factor binds specifically to the anti-sigma factor. They also found that the sigma factor and RNAP express mRNA, but transcription is interrupted when the anti-sigma factor is added. RT-PCR results show that the expression of the sigma factor varies depending on the carbon source. (Nataf et al., 2010a) (Figure 3)

GlyR3 and celC operon in *C. thermocellum*: Newcomb et al. have found that the *celC* operon is regulated by GlyR3, which is a LacI family protein. The celC operon contains the coding region for the *celC, glyR3* and *licA* genes, respectively. Those three genes share one promoter. (Figure 4)

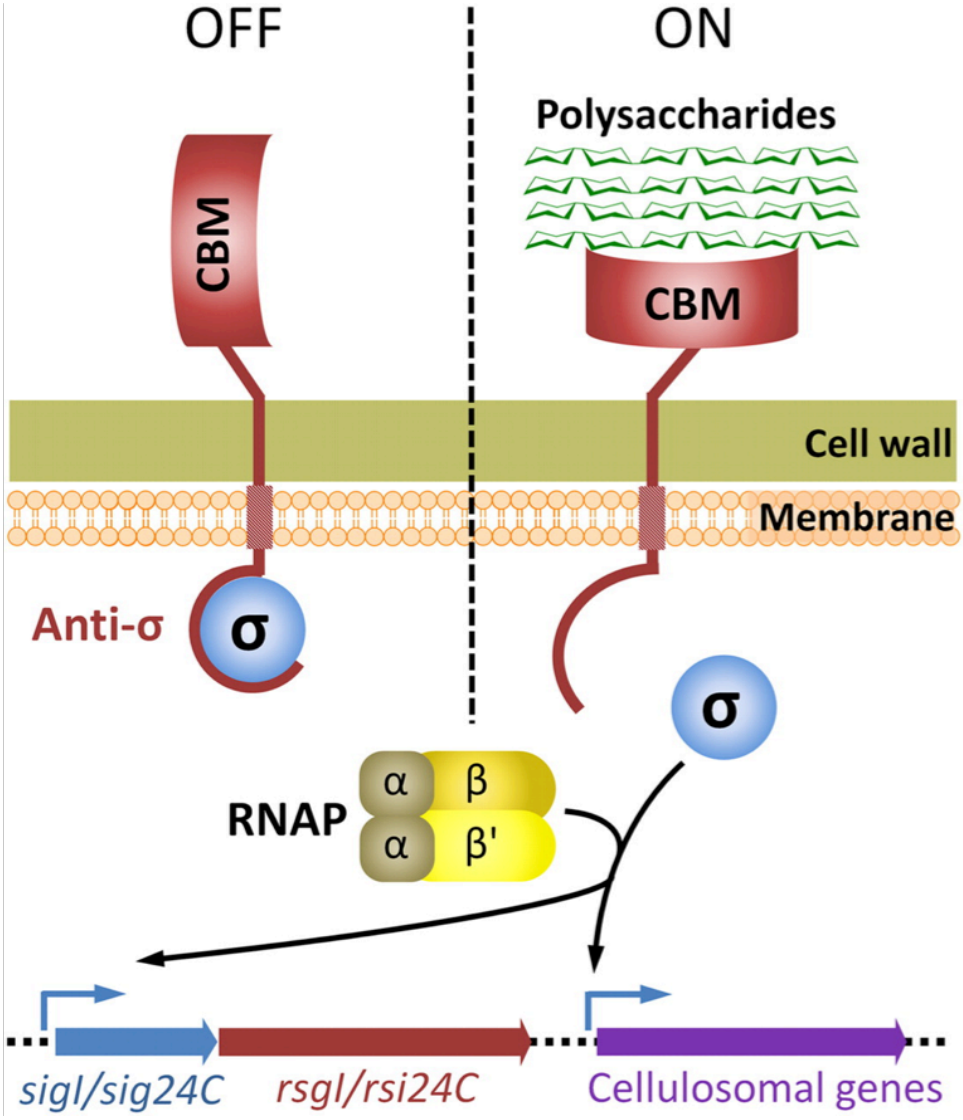


Figure 3 Mechanism for anti-sigma factor

(Nataf et al., 2010a)

The first gene produces CelC, a glycoside hydrolase (GH) family 5 protein that has beta-1-4-glucanase (cellulase) activity. CelC is a non-cellulosomal cellulase. *Clostridium thermocellum* has more than 100 glucanases including more than 70 cellulosomal cellulases.

The second gene produces GlyR3. GlyR3 is a LacI family regulatory protein that has Helix-Turn-Helix (HTH) motif to bind to DNA. The GlyR3 binds to operator of *celC* operon to suppress its own expression. If laminaribiose binds with GlyR3 it loses its ability to bind to the celC operon. Newcomb et al. (Newcomb et al., 2007a) found binding between celC operon and GlyR3 by DNA footprinting and EMSA. Addition of laminaribiose to the EMSA assay negatively affected the binding of GlyR3 to the *celC* operon. Also, gene expression patterns predicted by the EMSA experiments were verified *in vitro* by RT-PCR.

The last gene produces LicA. The LicA has a beta-1-3-glucanase activity. LicA breaks beta-1-3 glucan bonds. Laminaribiose is transported into the cell by ABC transporters (Nataf et al., 2009a) where it binds to GlyR3 to induce expression of the celC operon.

Recent research shows *manB* and *celT,* located downstream of *celC* operon, are co-regulated with celC. However, the mechanism of co-regulation is unknown. This dissertation tests the hypothesis that the GlyR3 protein is involved in the regulation of *manB* and *celT*. (Newcomb, Chen, & Wu, 2007b; Newcomb, Millen, Chen, & Wu, 2011b)

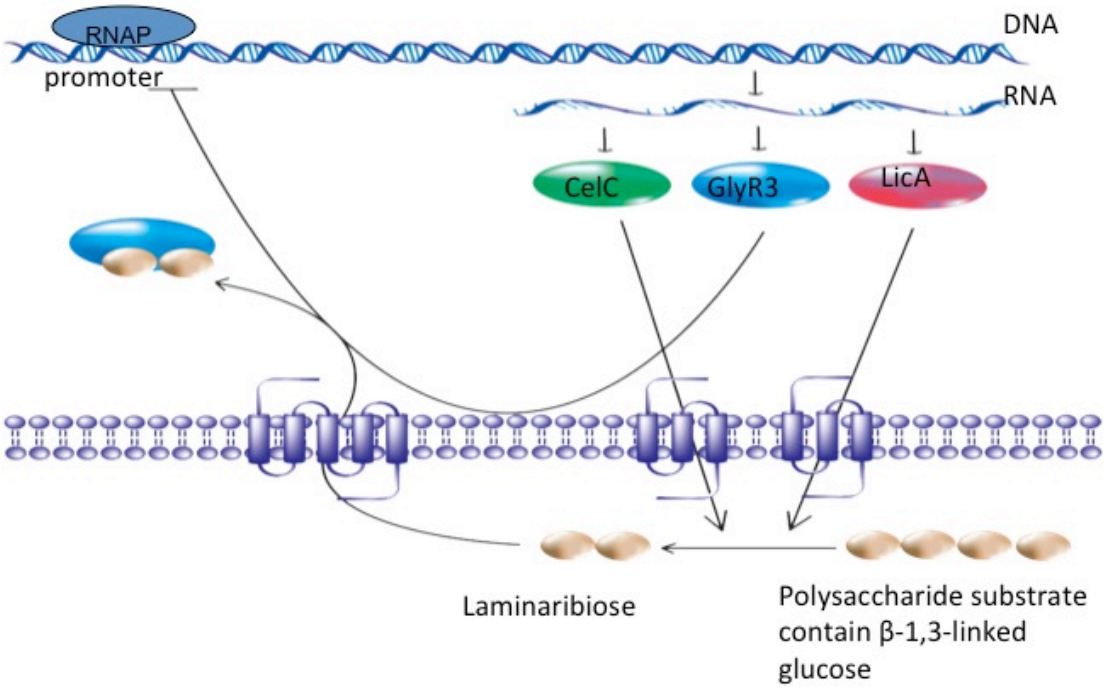


Figure 4. The celC Operon.

(Newcomb et al., 2007b)

Identification of transcriptional factor binding sites by computational methods

Multiple alignments: Regulatory proteins bind specific, conserved, DNA sequences. While these sequences are conserved, it is rare to find a perfect match to the consensus sequence. The variation of the DNA sequence from the consensus sequence results in the regulatory protein having different strengths of binding to various DNA sequences. The conservation of sequences makes it possible to identify potential DNA binding sequences by automated searching.

Position specific scoring matrix with information content: Information content is a method that applies information theory to quantify the similarity of a putative binding sequence to a group of known sequences. Information theory quantifies the information content of communication and was founded by Claude Shannon in 1948. The theory is used in various fields. The information content weighs the uncertainty of finding a given base at a given position by using the concept of entropy. The general formula for uncertainty is:

$$H \equiv -\sum_{i=1}^{\Omega} p_i \log_2 p_i$$

Where, H is the uncertainty, p_i is the i^{th} symbol's probability at a specific location in the DNA, Ω is number of symbols (Ω=4 for DNA).

Log base 2 was chosen because it represents bits of information communicated by a certain of number of symbols. For example, if you want to determine which of 2 boxes contains candy, you need 1 trial or question to find the answer ($\log_2 2=1$). But if there are 4 boxes, you need at least 2 questions ($\log_2 4=2$). If there are 8 boxes, you need 3 questions ($\log_2 8 =3$). So, log base 2 can represent the least number of questions needed to find the answer.

The formula is also used as information entropy. So, uncertainty and information entropy are synonyms (Jaynes, 1957). When the concept is applied to represent uncertainty of sequence information, it is reasonable to use uncertainty to describe its information content. In the case of searching protein-binding site on DNA, we can use the concept of 'before' and 'after' states. 'Before' state means uncertainty of maximum choices. 'After' state means uncertainty of choice after binding. So, the information content of the substrate can be described in terms of its before and after state.

$$R = \left[-\sum f \log_2 f\right] - \sum_{i=1}^{\Omega} p_i \log_2 p_i$$

Where, R is the information content, f is overall frequency of background. This equation can be rewritten to show individual information of each symbols of each position as:

$$R_i = [-\log_2 f] — \log_2 p_i$$

Where, R_i is the information content of each symbols of each position.

Schneider et al. (T. D. Schneider, Stormo, Gold, & Ehrenfeucht, 1986) assume that the overall frequency of each base in the genetic background has equal probability since the composition of the nucleic acid should not matter before the protein has made physical contact. Schneider et al. think overall frequency is close to 2 bits because they consider *E. coli* which has a GC content of 50.79%. Then, the 'before' state can be replaced by 2 (f=1/4).

$$R_i = 2 — \log_2 p_i$$

The sequence logo is a graphical representation of the information content of a conserved sequence determined using this equation (Crooks, Hon, Chandonia, & Brenner, 2004; T. D. Schneider & Stephens, 1990). This graph shows the information content from an aligned set of sequences for the LacI family protein binding site (CcpA) of *Bacillus subtilis* (Figure 5). We are using the CcpA binding site as an illustrative example because, as will be shown in Chapter 1, it has a great deal of similarity to the GlyR3 binding site in *C. thermocellum* (Choi, Klingeman, Brown, & Cox). The probabilities (p_i) are determined empirically from 44 sequences of known binding sites (Sierro, Makita, de Hoon, & Nakai, 2008). According to the information content, there is highly conserved sequence

positioned near the bases pair 11~24 in Figure 5. It tells us the consensus is TGAAAGCGCTTTCA.

However, Stormo et al. assume that overall frequency does not have equal probability in the biased genome. Stormo et al. think that if the frequency of bases across the genome is not 'equiprobable', the overall frequency can be important to be considered (Stormo, 1998). In this case, information content can be described as:

$$R_i = \log_2 \frac{p_i}{f}$$

In this dissertation, we use the equation from Stormo et al. because *C. thermocellum* has a GC content of 39.0%.

Table 1 shows a Position Specific Scoring Matrix (PSSM) (T. D. Schneider, 1997; T. D. Schneider et al., 1986) of the information content in the CcpA protein of *B. subtilis*. PSSM shows the score for each base at each position. We can use the PSSM to calculate the information content of any sequence by adding the score for a particular base at a particular position. For example, the score of TGCAATCGCTTACA is 1.67 + 1.84 + (-2.62) + 1.84 + 1.87 + (-2.04) + 1.81 + 1.84 + 0.63 + 1.59 + 1.08 + 0.38 + 1.51 + 1.47 = 12.86. The probability of a given sequence being a CcpA binding site increases with its PSSM score.

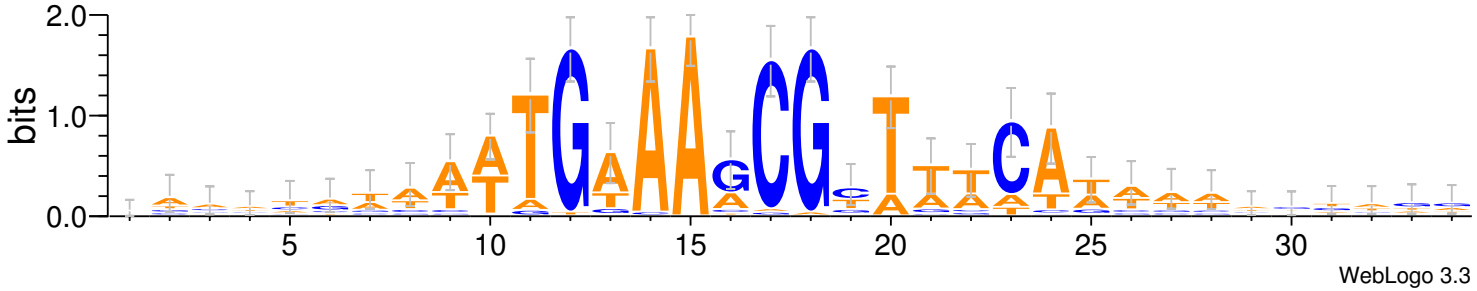


Figure 5 Web logo of the CcpA binding site in *B. subtilis.*

Table 1 Position Specific Scoring Matrix of CcpA binding motif in *B. subtilis.*

Base	Position													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	-1.30	-3.62	1.24	1.84	1.87	0.19	-2.62	-2.62	-1.62	-0.45	0.19	0.38	-0.81	1.47
G	-2.04	1.84	-1.30	-3.62	-3.62	1.13	-2.62	1.84	-0.62	-3.62	-1.30	-1.62	-3.62	-2.62
C	-3.62	-3.62	-2.62	-2.62	-3.62	-1.62	1.81	-3.62	0.63	-3.62	-2.04	-1.62	1.51	-2.04
T	1.67	-2.62	-0.04	-3.62	-3.62	-2.04	-3.62	-3.62	0.47	1.59	1.08	0.96	-1.30	-0.45
	T	G	A	A	A	G	C	G	C	T	T	T	C	A

Summary of chapter1

Determination of additional GlyR3 regulation sites: the *celC* operon is regulated by GlyR3, which is a LacI family protein (Newcomb et al., 2007a). CcpA is a well-characterized global regulator in *B. subtilis* that has a LacI motif and is known to regulate 44 genes. CcpA and GlyR3 have a high level of similarity and share a highly conserved helix-turn-helix domain that is known as a DNA binding domain (Figure 6) (Schumacher, Seidel, Hillen, & Brennan, 2006). Also, protein sequence alignment shows that those proteins are highly similar in their amino acid sequence within this domain.

So we use the information content of the CcpA binding sites in *B. subtilis* to search for additional putative GlyR3 binding sites in *C. thermocellum*. The information content characterizes the conserved information at each position as was described in the background section. The average information content score from the PSSM is 15.34 for CcpA sites in *B. subtilis* and 42 of the 44 sequences have information content scores greater than 12. Experimental verification is necessary to sort true positives from false positives, because there are a total of 801 sequences in *B. subtilis* which score greater than 12.

The probability of finding a true binding site can be increased if we add more conditions. We searched sequences near the *celC* operon and found potential binding sites. We found a sequence in the coding region of *manB* with a high score and another on an upstream region of *celT* with a slightly lower score. These binding sites were considered to be particularly good candidates for

additional investigation since the *manB* - *celT* cluster was recently shown to be co-regulated with *celC* operon (Newcomb et al., 2011b).

For these reasons, we hypothesize that GlyR3 is involved in regulation of *manB* and *celT* located downstream of the *celC* operon. EMSA test is performed to see evidence of direct binding of GlyR3. The EMSA results shows that the putative binding site in *manB* binds to GlyR3 but its affinity is weaker than *celC*. The putative binding site in *celT* dose not show binding. Additional EMSA tests were performed with addition of laminaribiose. The result show that GlyR3 lost its ability to bind to the celC sequence when laminaribiose is added, consistent with the behavior observed earlier by Newcomb(Newcomb et al., 2007a). In contrast, laminaribiose has no effect on the ability of GlyR3 to bind to the manB sequence. Next, we measured gene expression change upon addition of laminaribiose both *in vivo* and *in vitro*. The expression profile in *celC* is induced after adding laminaribiose and its change was proportional of concentration of laminaribiose. In *manB*, the expression was repressed in the presence of laminaribiose but the change in expression was not proportional to the concentration of laminaribiose. The same expression pattern was observed in an i*n vitro* transcription assay. The expression is repressed with adding GlyR3 in both *celC* and *manB*. In *celC*, the repression is relieved when laminaribiose is added. In *manB*, the repression remained.

A model for the regulation of the *celC* and *manB* operons is suggested based on the experimental results. When laminaribiose is not present, the *celC*

operon is repressed by low levels of GlyR3, while *manB* is able to be expressed. If laminaribiose is added, *celC* expression is activated. Then high level of GlyR3 represses *manB* expression.

Summary of Chapter 2

Identification of celC-type and manB-type GlyR3-binding motifs in *C. thermocellum*: The presence of many additional sites with relatively high CcpA PSSM scores in the *C. thermocellum* genome suggests that it is possible that GlyR3 regulates additional operons. Also, RNA sequencing data shows that 659 genes change expression under conditions in which GlyR3 is overexpressed by adding laminaribiose.

So, we extend the GlyR3 binding search to the whole genome in *C. thermocellum* to find additional GlyR3 binding sites. We found that 126 operons have putative GlyR3 binding site with a score over 12 based on the CcpA motif. Experimental confirmation is necessary, which can be done by EMSA and IVT. Eleven candidates were tested by EMSA for evidence of direct binding of GlyR3. Seven of them show binding to GlyR3 but four genes did not. Two of the candidates were further investigated by adding laminaribiose in EMSA. Cthe_0391 is a gene that is up-regulated in the presence of laminaribiose. EMSA testing reveals that that is binds GlyR3 but that the binding is released to a large extent in the presence of laminaribiose.

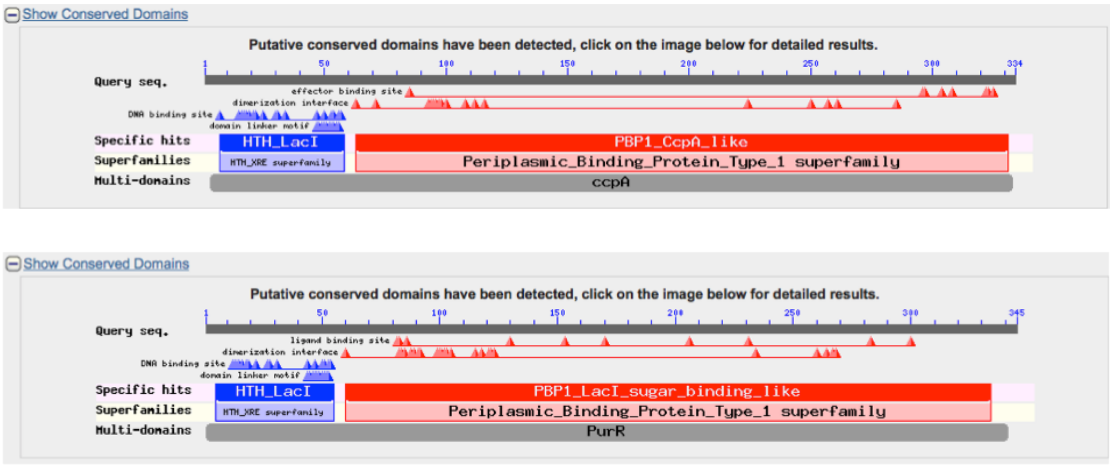


Figure 6 Conserved Domain Search.

Similarity of CcpA and GlyR3. Both CcpA and GlyR3 have a helix-turn-helix LacI domain

Cthe_0480 is down-regulated in the presence of laminaribiose. EMSA testing reveals that it binds GlyR3 both with and without laminaribiose. *In vitro* transcription assay is performed with those two candidates. Both candidates were repressed upon addition of GlyR3 but only 0391 showed a reversal of the of repression upon addition of laminaribiose. Across all experiments, Cthe_0391 behaved in a similar manner as celC, while Cthe_0480 behaved similarly to manB.

We hypothesized that the differences in binding behavior of these two groups of genes many be related to the interactions of GlyR3 with two distinct binding motifs. Based on this hypothesis, we created a celC-type motif based on the GlyR3 binding sequences of celC, Cthe_0391 and three additional genes that increased expression in the presence of laminaribiose and a manB-type motif based on the sequences of manB, Cthe_480 and two additional genes down regulated in the presence of laminaribiose. We then searched the *C. thermocellum* genome using each of these motifs and found that operons associated with celC-type binding motifs were significantly enriched in up regulated operons and that operons associated with manB-type motifs were significantly enriched in down-regulated operons. Genes associated with carbohydrate transport and metabolism were associated with the celC-type motif. Genes associated with carbohydrate transport and metabolism; chromatin structure and dynamics; DNA replication, recombination and repair; energy

production and conversion; and signal transduction mechanisms are associated with manB-type motif.

CHAPTER I
THE LACI FAMILY PROTEIN GLYR3 CO-REGULATES THE *CELC* OPERON AND *MANB* IN *CLOSTRIDUM THERMOCELLUM*

A version of this chapter will be submitted for publication by Jinlyung Choi and Chris Cox:

The manuscript will be submitted to Biotechnology for Biofuels.

Jinlyung Choi designed and conducted the experiments, analyzed the data, and wrote the paper.

1.1 Abstract

Clostridium thermocellum utilizes a wide variety of free and cellulosomal cellulases and accessory enzymes to hydrolyze polysaccharides present in complex substrates. To date only a few studies have unveiled the details by which the expression of these cellulases are regulated. Recent studies have described the auto regulation of the celC operon (Newcomb et al., 2007a) and determined that the celC-glyR3-licA gene cluster and nearby manB-celT gene cluster are co-transcribed as polycistronic mRNA (Newcomb, Millen, Chen, & Wu, 2011a).

In this paper we first identify putative GlyR3 binding sites within or just upstream of the coding regions of *manB* and *celT*. Using an electrophoretic mobility sift assay (EMSA) we determined that the putative *manB* site binds GlyR3 more weakly than the *celC* site. Neither the putative *celT* site nor random DNA significantly binds GlyR3. While laminaribiose interfered with GlyR3 binding to the *celC* binding site, binding to the *manB* site was unaffected. Expression of *manB* was repressed in the presence of laminaribiose both in vivo and in vitro assay.

Together these results reveal a mechanism by which *manB* is expressed at low concentrations of GlyR3 but repressed at high concentrations. In this way *C. thermocellum* is able to co-regulate both the *celC* and *manB* gene clusters in response to the availability of β-1,3 polysaccharides in its environment.

1.2 Background

Clostridium thermocellum is an anaerobic, thermophilic, Gram-positive bacterium that has a highly efficient cellulolytic system (Lynd, Weimer, van Zyl, & Pretorius, 2002a). This bacterium is considered to be a model organism for biofuels processing since it combines cellulolytic and ethanologenic abilities (Carere, Sparling, Cicek, & Levin, 2008; Demain, Newcomb, & Wu, 2005; Lynd et al., 2002a; Maki, Leung, & Qin, 2009; Olson & Lynd, 2012; Raman, McKeown, Rodriguez, Brown, & Mielenz, 2011). Cellulolytic activity is conferred by a combination of free glycoside hydrolases and an extracellular multi-enzyme cellulase complex called the cellulosome (Bayer, Belaich, Shoham, & Lamed, 2004a; Fontes & Gilbert, 2010; Garciamartinez, Shinmyo, Madia, & Demain, 1980; Gilbert, 2007; Schwarz, 2001; Sheehan & Himmel, 1999). *C. thermocellum* ATCC 27405 is the reference strain. Strains YS and AD2 were used in many of the key studies which developed the cellulosome concept and have recently been sequenced (Brown et al., 2012). An efficient transformation methodology has been developed for strain DSM 1313 (Tripathi et al., 2010) facilitating the development of an engineered strain capable of high ethanol titter (Argyros et al., 2011).

C. thermocellum employs more than 100 genes for biomass degradation, including more than 70 genes that encode for various cellulosomal enzymes (Newcomb et al., 2007a). The cellulosome has a core, scaffold protein called CipA that binds to the surface of the bacterial cell, to the catalytic subunits, and to the carbohydrate-binding module (CBM) (Newcomb et al., 2011a). Various CBM and catalytic subunits may be deployed to provide cellulolytic activity specific to various biomass substrates (Bahari et al., 2011a). While many studies have described the structural and catalytic activity of the cellulosome and free cellulases, relatively few investigations have focused on the regulation of these genes; the most significant of these are reviewed below.

Recently it was determined that many cellulosomal genes are regulated by a common mechanism involving the s^I alternate transcription factor, which binds to the core RNA polymerase to form a holoenzyme capable of transcribing these genes (Nataf et al., 2010b). In the absence of polysaccharides, SigI is inactivated via binding to the anti-sigma factor N-terminal domain of the trans-membrane protein RsgI. The conformation of RsgI changes upon binding of a target extracellular polysaccharide to the C-terminal CBM of the RsgI protein, thereby releasing SigI to the cytoplasm of the cell and up regulating SigI-regulated genes, including *sigI* and many cellulosomal genes. Various SigI-RsgI proteins are activated by specific polysaccharides, thereby providing specificity in regulation of cellulosomal genes. (Nataf et al., 2009b).

In contrast, the *celC* operon, containing the *celC*, *glyR3*, and *licA* genes, is regulated by a different mechanism involving the LacI family protein GlyR3, which negatively auto-regulates the operon by binding to the *celC* promoter region to repress its expression (Newcomb et al., 2007a). The repression of the operon is relieved in the presence of laminaribiose, which interferes with GlyR3 binding to the promoter. Regulation of the *celC* operon is perhaps the most well characterized of the non-cellulosomal cellulases in *C. thermocellum*. CelC is a non-cellulosomal endoglucanase affiliated to the glycoside hydrolase family 5, which is one of the largest of the glycoside hydrolase families. LicA is an endo-1,3-beta-D-glucosidase. Recently it has been shown that the nearby *manB-celT* gene cluster is co-regulated with *celC* by an unknown mechanism (Newcomb et al., 2011a). It was also shown that the *manB-celT* gene cluster was transcribed as a single operon (Newcomb et al., 2011a). ManB is a cellulosomal family 26 glycoside hydrolase and CelT is a cellulosomal family 9 endoglucanase (Halstead, Vercoe, Gilbert, Davidson, & Hazlewood, 1999; Kurokawa et al., 2002).

In this paper, we demonstrate an inverse relationship between *glyR3* and *manB* gene expression and identify a binding site of GlyR3 within the coding region of *manB*, thereby allowing us to infer a GlyR3-dependent regulatory mechanism for *manB* in *C. thermocellum*. In other gram-positive organisms, LacI family proteins similar to GlyR3 are known to repress numerous carbon metabolism pathways (Henkin, 1996). This result opens the possibility of GlyR3

playing a larger role in regulating *C. thermocellum* cellulolytic activity than previously known.

1.3 Results

1.3.1 Protein and DNA sequences suggest similarities in DNA binding between GlyR3 in C. Thermocellum and CcpA in B. subtilis

CcpA is a LacI family protein in *Bacillus subtilis* that is known to regulate at least 44 different operons (Sierro et al., 2008). We aligned the helix-turn-helix domains of CcpA and GlyR3 to determine the similarity of the their DNA binding domains. The two proteins show a high degree of similarity over their first 60 residues (Figure 7). Of the 18 residues in direct contact with DNA, 14 were exact matches. We also observed that the GlyR3 binding site (TGAACGCGCGTACA) in the *celC* operon was similar to the consensus CcpA binding site in *B. subtilis* (TGNAANCGNWNNCW). These two observations led us to hypothesize that CcpA could be used as a model to identify additional GlyR3 binding sites in *C. thermocellum.*

1.3.2 Putative GlyR3 binding sites associated with manB and celT are identified

The DNA sequences of the 44 known CcpA binding sites in *B. subtilis* (Sierro et al., 2008) were used to construct a position specific scoring matrix (PSSM) to search for additional potential GlyR3 binding sites in *C. thermocellum*.

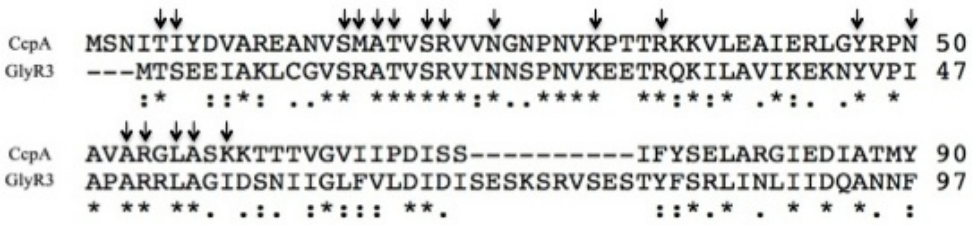


Figure 7 Amino acid sequence alignment

Clustal 2.1 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) multiple sequence alignment showing high similarity of the first 60 amino acid sequences of CcpA in *B. subtilis* and GlyR3 in *C. thermocellum.* There are 14 matching out of a total of 18 amino acid(Schumacher et al., 2006). (*****(asterisk) indicates positions which have a single, fully conserved residue; **:** (colon) indicates conservation between groups of strongly similar properties, scoring > 0.5 in the Gonnet PAM 250 matrix; **.** (period) indicates conservation between groups of weakly similar properties, scoring =< 0.5 in the Gonnet PAM 250 matrix; arrow indicates amino acid in direct contact with the DNA) (Larkin et al., 2007)

Using the *B. subtilis*-derived PSSM provided information about the relative importance of each base within the sequence that would not have been available had we searched using the *celC* GlyR3 site alone. The performance of the PSSM in identifying CcpA binding sites in *B. subtilis* is given by the receiver operating characteristic (ROC) curve, which is a graphical illustration of the performance of a threshold in a binary classifier system (Hanley & McNeil, 1982). The ROC curve demonstrates that with a threshold value of 12 the PSSM is able to identify a high fraction of CcpA sites in *B. subtilis* (high sensitivity) while maintaining a relatively low false positive rate (1-specicifity) (Figure 7). The arrow in Figure 7 corresponds to a 1/10 true positive to false positive ratio and is shown for reference. The PSSM score of the GlyR3 binding site of *celC* is 9.18. Which would correspond to a 1/50 true positive to false positive ratio. While the celC GlyR3 binding sequence has some deviation from the CcpA motif, there are still significant similarities such that the PSSM derived from CcpA may be useful in identifying potential GlyR3 binding sites in *C. thermocellum.* Putative GlyR3 binding sites in the vicinity of the celC operon were identified by scanning the *C. thermocellum* genome using the PSSM for CcpA and are listed in Table 1. A putative GlyR3 binding site for *manB* with a PSSM score of 14.62 was identified. Since the score of the *manB* site was greater than the score of the *celC* (9.18), this was a promising site for GlyR3 binding. In addition a putative GlyR3 binding site for *celT* was also identified. While the *celT* PSSM score of 5.78 was

significantly lower than that of *celC* and *manB*, it was significantly greater than the average PSSM score for a random location in the overall genome (-14.08).

1.3.3 GlyR3 binds to the putative binding site in the manB coding region

Electrophoretic Mobility Shift Assay (EMSA) was used to further investigate protein-DNA interactions in the presence and absence of laminaribiose for the GlyR3 binding regions of *celC*, *manB*, and *celT*. The binding region *celC* was included to show consistency with the binding behavior previously reported in reference (Newcomb et al., 2007a). In the absence of laminaribiose, addition of GlyR3 results in a strong shift in the *celC* band, a partial shift in the location of the *manB* band, and no shifting of the *celT* or random DNA bands (Figure 9).

The top band in lane 2 of Figure 2b can be attributed to high-molecular-weight DNA-GlyR3 aggregates (Fried & Crothers, 1981). The GlyR3-induced shifts of the *celC* and *manB* bands were reversed upon addition of competitor DNA (unlabeled 18-mer), confirming that the putative binding sites were responsible for the GlyR3 binding.

The strength of DNA binding by GlyR3 was further investigated using a titration test in which the GlyR3 concentration was increased while keeping the DNA concentration constant (Figure 10). It was observed that GlyR3 interactions with the *celC* binding domain were insignificant at levels less than or equal to 15 ng of protein but that all DNA was bound by GlyR3 at levels equal to or greater

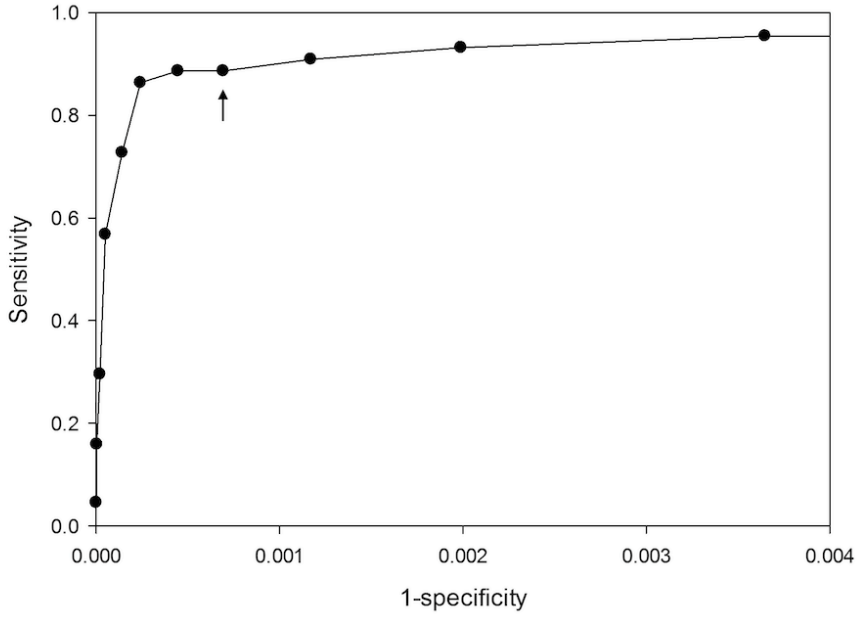


Figure 8 ROC curve

Binding search based on CcpA motif in *B. subtilis* genome. The plot is scaled to enlarge high scored area. Eleven dots on the plot indicate bins of score 19 to 9 left to right. Vertical arrow indicates a 1/10 true to false positive ratio

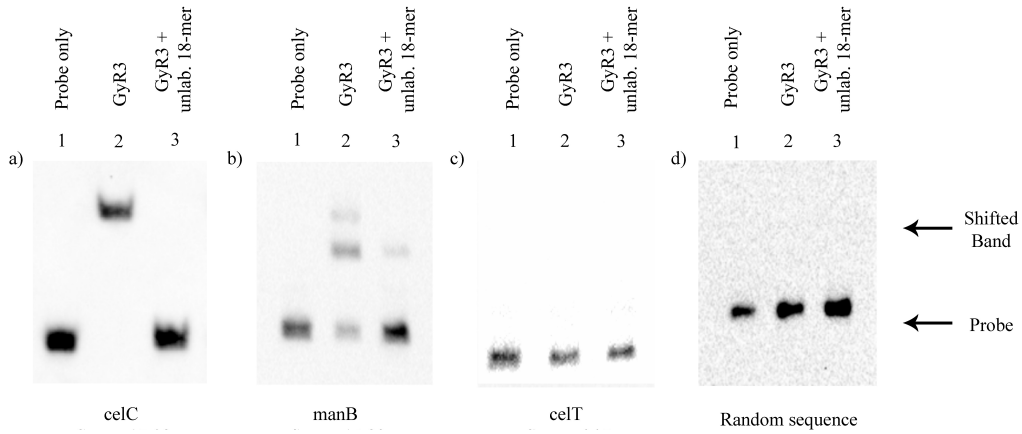


Figure 9 GlyR3 - DNA binding

Electrophoretic Mobility Shift Assay to assess binding of GlyR3 to putative DNA binding sites for *celC*, *manB*, *celT*, and a random sequence of DNA. a) Lane1 *celC* (0.2 ng (0.3 nM), Table 2. 3-4); Lane2: *celC*+GlyR3 (150 ng (387 nM)); Lane3: *celC*+GlyR3+ unlabeled 18-mer (500 ng (2.14 mM), Table 2. 11-12). b) Lane1: *manB* (0.2 ng (0.18 nM), Table 2. 5-6); Lane2: *manB*+GlyR3 (150 ng (387 nM)); Lane3: *manB*+GlyR3+ unlabeled 18-mer (500 ng (2.14 mM), Table 2. 13-14). c) Lane1: *celT* (0.2 ng (0.19 nM), Table 2. 7-8); Lane2: *celT*+GlyR3 (150 ng (387 nM)); Lane3. *celT*+GlyR3+ unlabeled 18-mer (500 ng (2.14 mM), Table 2. 15-16), d) Lane1: Random DNA (0.2g (0.2nM), Table 2. 9-10); Lane2: DNA+GlyR3 (150 ng (387nM)); Lane3: DNA+GlyR3+ unlabeled 18-mer (500ng (2.14 mM), Table 2. 11-12)

than 30 ng. In contrast, levels equal to or greater than 60 ng of GlyR3 were needed to achieve significant binding of the *manB* binding domain. The band was partially shifted upon addition of 60 ng and further shifted at 140 ng of GlyR3. Overall the data in Figure 11 suggest that the *celC* binding domain has a higher affinity for GlyR3 than the *manB* binding domain. Addition of 35 and 70 μg of laminaribiose was shown to relieve GlyR3 repression in a dose-dependent manner, consistent with previous reports (Newcomb et al., 2007a) (Figure 11). In contrast, GlyR3 binding to the *manB* binding site appeared to be unaffected by laminaribiose addition. Therefore the effect of laminaribiose on GlyR3 binding to DNA was different for the two binding sites, which may affect the way in which laminaribiose controls expression of the regulated genes.

1.3.4 In vivo expression of manB is repressed in the presence of laminaribiose

We determined the effect of adding laminaribiose on the expression of *celC*, *manB* and *celT* in *C. thermocellum* using qRT-PCR. *C. thermocellum* was incubated at 60°C for one hour after adding different concentrations of laminaribiose to each vial. The data was normalized to the housekeeping gene *recA*. Gene expression changes were determined by the comparative CT method using a control sample to which no laminaribiose was added (Schmittgen & Livak, 2008). As shown in Figure 12a, transcription of *celC* increased with laminaribiose concentration. All samples showed an increase in gene expression compared to samples without laminaribiose. In contrast, *manB* was repressed by

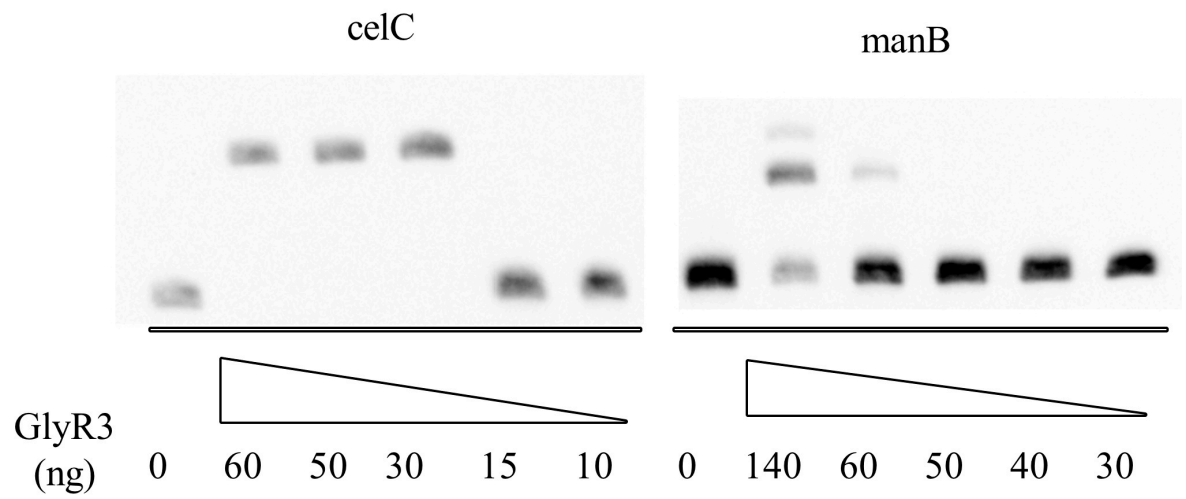


Figure 10 Titration of GlyR3

EMSA of *celC* and *manB* binding sites as a function of GlyR3 level. DNA loading was 0.2 ng (celC: 0.3 nM, manB: 0.18 nM). GlyR3: 140 ng (361 nM), 60 ng (155 nM), 50 ng (129 nM), 40 ng (103 nM), 30 ng (77 nM), 15 ng (39 nM), 10 ng (26 nM).

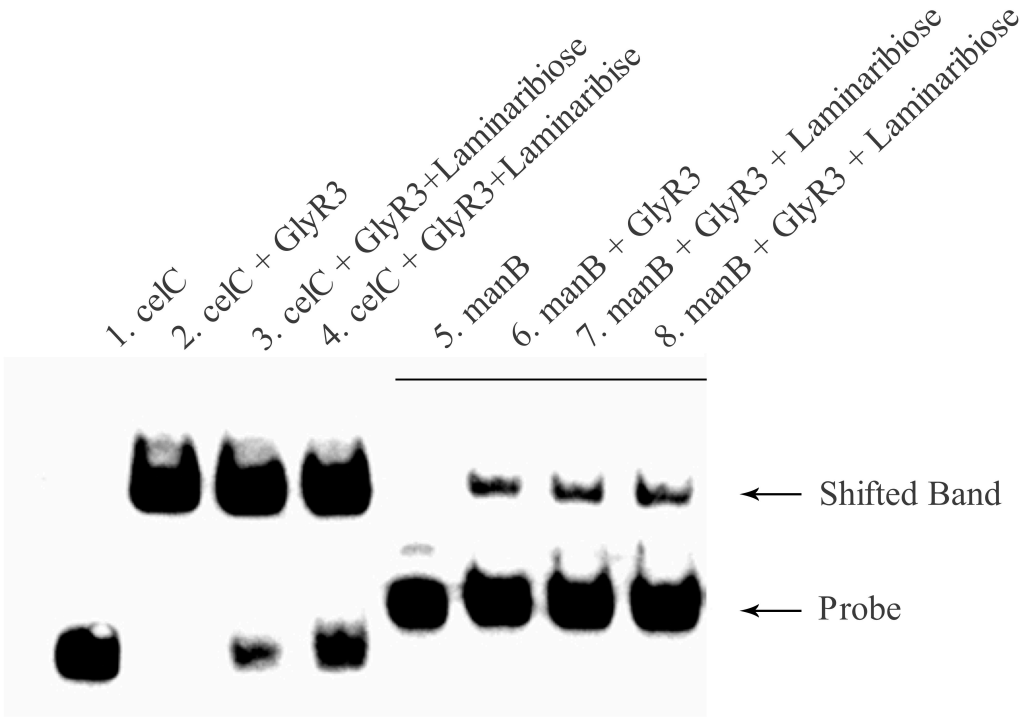


Figure 11 Laminaribiose effect of GlyR3

EMSA to determine the effect of laminaribiose addition on GlyR3 binding to DNA binding sites. Lane1: *celC* (0.2 ng (0.3 nM)); Lane2: *celC*+GlyR3 (150 ng (387 nM)); Lane3: *celC*+GlyR3+laminaribiose (35 μg (50 mM)); Lane4: *celC*+GlyR3+laminaribiose (70 μg (100 mM)); Lane5: *manB* (0.2 ng (0.18 nM)); Lane6: *manB*+GlyR3(150 ng (387 nM)); Lane7: *manB*+GlyR3+laminaribiose (35 μg (50 mM)); Lane8: *manB*+GlyR3+laminaibiose (70 μg (100 mM))

laminaribiose in a non-dose-dependent manner (Figure 12 b). All samples show a decrease in *manB* gene expression compared to samples without laminaribiose. Expression of *celT* was unaffected by laminaribiose addition (Figure 12 c).

1.3.5 Transcriptional start sites of celC and manB identified by mRNA sequencing

We used RNA sequencing to identify transcriptional start sites for the *celC* and *manB* gene clusters (Figure 13a and b). The transcriptional start site for *celC* approximately coincides with the GlyR3 binding site. Furthermore, we were able to identify candidate -10 and -35 sigma A promoter sequences with high fidelity to the consensus sequence just upstream of the transcription start sites for both the *celC* and *manB* (Figure 13a, b, and c).

The RNA-seq data also revealed an interesting feature of the *manB-celT* expression profile (Figure 13 d), in that there was a strong increase in expression at the beginning of the *celT* gene, suggesting the presence of a transcription initiation site between the two genes. However, we were unable to identify a nearby upstream sigma A promoter site for *celT*.

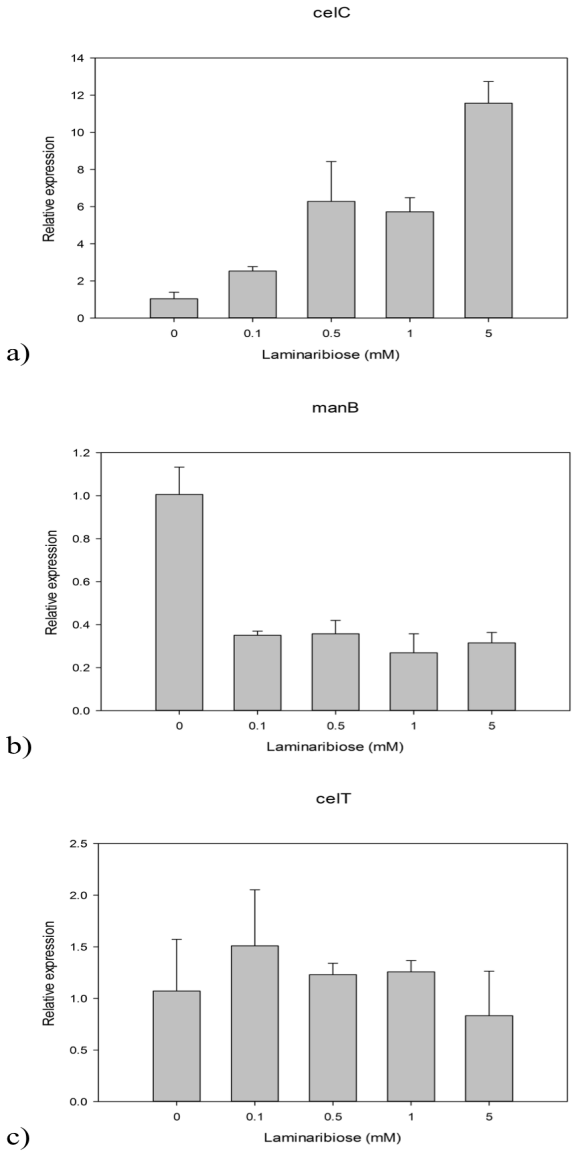


Figure 12 in vivo expression

Relative gene expression as a function of laminaribiose concentration as determined by quantitative RT-PCR. a) *celC*; b) *manB*; c) *celT*

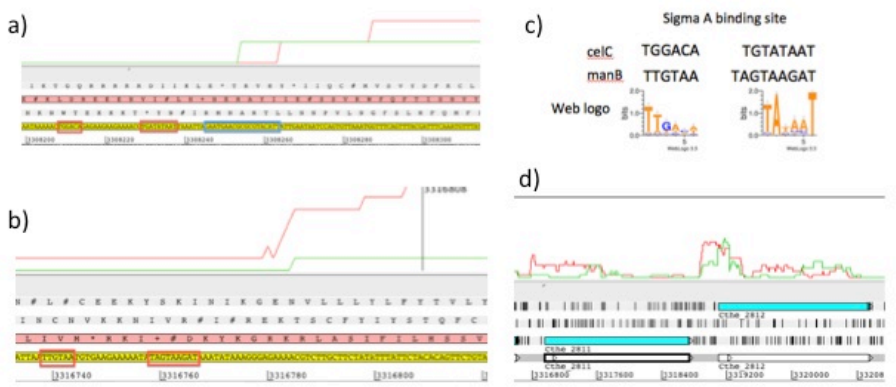
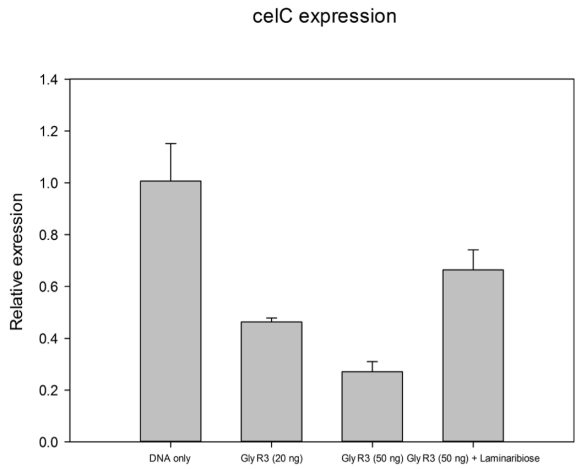


Figure 13 Transcription start site and binding site

a) celC mRNA expression profile (Red box: putative sigma A binding site, Blue box: GlyR3 binding site) (Red: Forward, Green: Reverse) b) *manB* mRNA expression profile (Red box : putative sigma A binding site) (Red: Forward, Green: Reverse) c) Putative sequences of sigma A promoters of celC and *manB* in comparison to the web logo of the consensus sequences of the -10 (right) and -35 (left) promoter regions. d) Expression profile over the *manB* (Chte_2811) and *celT* (Cthe_2812) gene cluster. (Red: Forward, Green: Reverse)

a)



b)

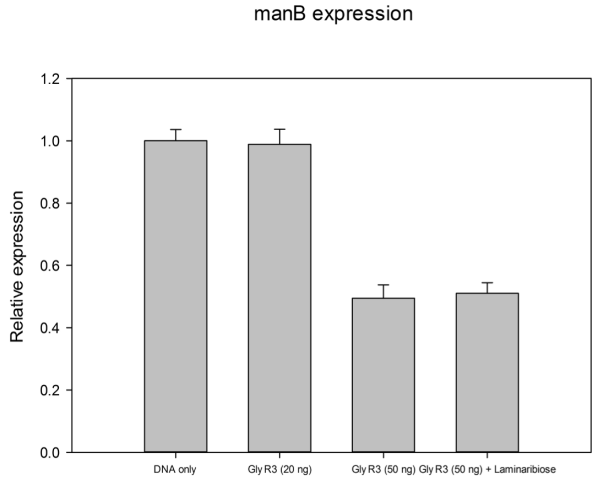


Figure 14 in vitro transcription assay

In vitro Relative gene expression as a function of laminaribiose concentration as determined by quantitative RT-PCR. a) *celC* (DNA: 2.65 nM, GlyR3: 20 ng (51.6 nM), 50 ng (128.9 nM), Laminaribiose: 50 mM); b) *manB* (DNA: 2.05 nM, GlyR3: 20 ng (51.6 nM), 50 ng (128.9 nM), Laminaribiose: 50 mM);

1.3.6 The role of GlyR3 in Repressing manB expression is confirmed by In Vitro Transcription Assay

As shown in Figure 14A, expression of *celC* was repressed by GlyR3 in a dose-dependent manner (p-value = 0.00294 for 20 ng GlyR3 vs. DNA only and p-value = 0.0013 for 50 ng vs. 20 ng GlyR3). Upon addition of laminaribiose repression by GlyR3 was relieved (p-value = 0.00141 between GlyR3 50 ng and GlyR3 50 ng + Laminaribiose). This result showed a same pattern from Newcomb et al. The gene *manB* was not repressed upon addition of 20 ng of GlyR3 (p=0.74) but was repressed at 50 ng of GlyR3 (p=0.0001). However, the repression of *manB* was not affected by laminaribiose (p=0.64) at a dosage of 50 ng of GlyR3. This pattern is consistent with the results of the EMSA and *in vivo* expression assays.

1.4 Discussion

Our results suggest an extended model of the *celC* regulon that includes regulation of *manB* via GlyR3 (Figure 15). In the model proposed by Newcomb et al., expression of the *celC* operon is auto-repressed in the absence of laminaribiose due to binding of GlyR3 to the *celC* promoter (Newcomb et al., 2007a). Our EMSA results show that binding of GlyR3 to the *manB* binding site is relatively weak compared to GlyR3 binding to the *celC* binding site. Therefore, the low concentration of GlyR3 that is likely to occur when the *celC* operon is autorepressed may be insufficient to repress the expression of *manB*. Indeed, *manB* expression was observed in the absence of laminaribiose in the gene

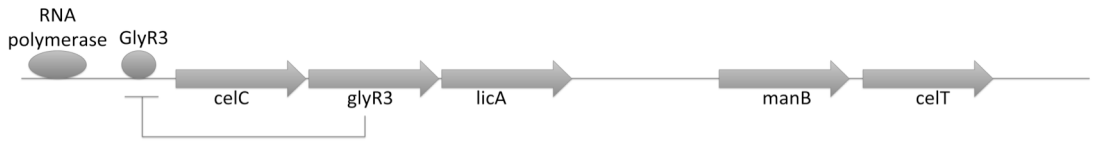
expression results presented here and earlier by Newcomb et al (Newcomb et al., 2011a).

Laminaribiose relieves the repression of GlyR3 according to the *celC* model of Newcomb (Newcomb et al., 2007a). Under these conditions, GlyR3 may be expressed at sufficiently high levels to bind to the *manB* binding site, thereby blocking expression of *manB*. Gene expression data presented here confirms that *manB* expression decreases in the presence of laminaribiose (Figure 12b). Consistent with this observation, Newcomb (Newcomb et al., 2011a) also observed a decrease in *manB* expression when grown on laminarin in comparison to cellulose. In contrast to GlyR3 binding to the *celC* binding site, GlyR3 binding to the *manB* binding site is mostly unaffected by laminaribiose according to the gene expression data (Figure 12b and Figure 14b) and EMSA experiments (Figure 11).

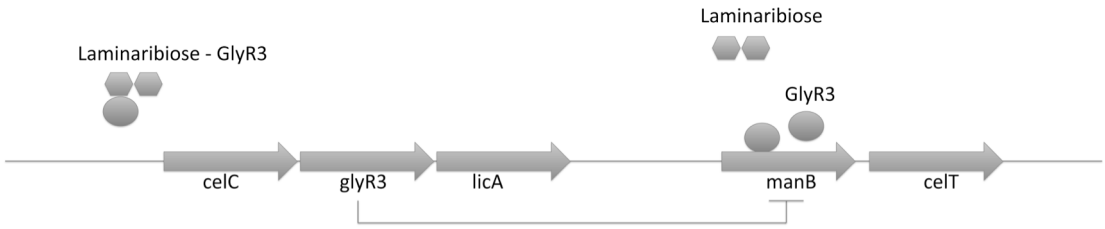
The mechanism by which laminaribiose would decrease binding at the *celC* site, but not affect binding at the *manB* site is unknown, but could be related to different conformations of the GlyR3 protein when it bounds to the two different DNA sequences. Overall, this model allows for repression of *manB* and expression of the *celC* operon to be under control of laminaribiose. This contrary behavior is critically dependent both upon the weaker relative binding of GlyR3 to the *manB* binding site and the apparent lack of effect of laminaribiose on the GlyR3-*manB*-binding-site complex.

We were able to identify DNA sequences consistent with the σ_A promoter directly upstream of the transcriptional start sites identified by mRNA sequencing. The presence of a σ_A promoter not only lends confidence to the location of the transcriptional start site, but it also provides mechanisms for expression of these gene clusters under “housekeeping” conditions. The transcriptional start sites identified in this work differ from those reported by Newcomb (Newcomb et al., 2011a). These differences possibly represent alternative transcripts regulated by alternative mechanisms that may be operative under the different experimental conditions.

Newcomb (Newcomb et al., 2011a) presented evidence that *manB* and *celT* form an operon which produces polycistronic mRNA when transcribed. Our gene expression data showed that while *manB* expression was regulated by laminaribiose, *celT* expression was unaffected (Figure 14c). The mRNA sequence data in Figure 6d, indicates an increase in expression near the beginning of the *celT* coding region. These lines of evidence suggest a possible alternate transcription initiation site for *celT*, independent of *manB*. Our bioinformatics analysis identified two potential weak GlyR3 binding sites located just upstream of the *celT* coding region (Figure 16a). EMSA showed no evidence of GlyR3 binding by the more upstream of the two sites (Figure 9c), and showed only very weak binding by the downstream site (data not shown). Expression of *celT* did not appear to be affected by the presence of laminaribiose further suggesting that GlyR3 binding was not functional at these sites.



(a) Without laminaribiose



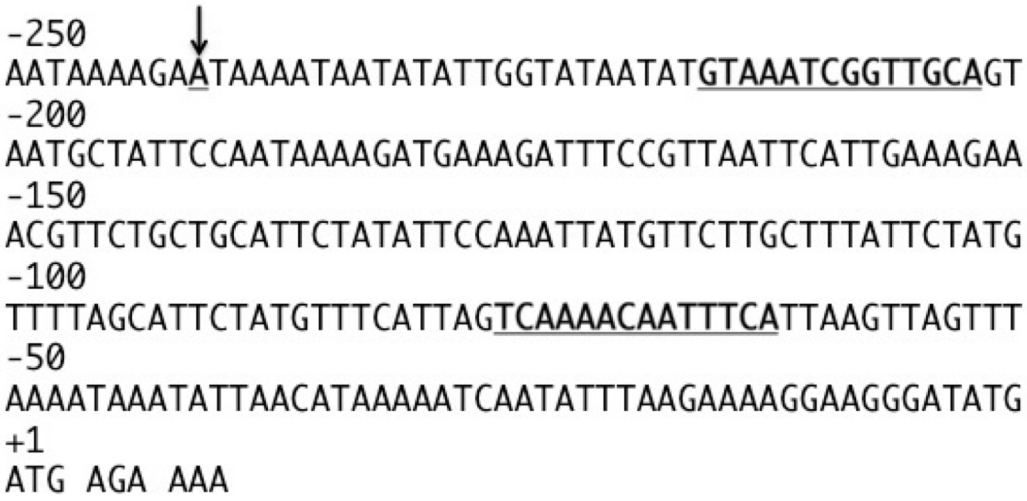
(b) With laminaribiose

Figure 15 model of GlyR3 regulation

Expanded model of *celC* regulon. a) In the absence of laminaribiose, GlyR3 autosuppresses the expression of the *celC* operon, resulting in relatively low GlyR3 concentrations, thereby allowing expression of *manB*. b) In the presence of laminaribiose, repression of the *celC* operon is relieved resulting in high GlyR3 concentrations and repression of manB.

However, these two sequences characterized by relatively low PSSM scores (in the range of 6 or 7), could mutate into sequences characterized by relatively high PSSM scores (in the range of 10 to 15) through one or two hypothetical single-nucleotide polymorphisms (SNP) (Figure 16b). It is possible that these GlyR3 binding sites may have played a role in *celT* regulation in the evolutionary history of the organism.

As discussed above, we identified a σ_A promoter region that appears to regulate *manB* expression, at least under certain conditions. In addition to this regulatory mechanism, the *manB* operon is among those recently reported to be regulated by the σ/anti-σ factor mechanism (Nataf et al., 2010b). Specifically, *manB* is regulated by the σ^{I6} factor. The N-terminal sensing domain of the corresponding RsigI6 transmembrane protein is sensitive to xylene and cellulose. This mechanism would seem to up regulate the *manB* operon in the presence of a broad spectrum of biomass substrates. Regulation by GlyR3 may be superimposed on the σ/anti-σ factor mechanism. The GlyR3 mechanism tends to down regulate the *manB* operon in the presence of β -1,3 polysaccharides, while at the same time up regulating *celC* and *licA*. ManB is a cellulosomal cellulase with β -1,4 hydrolytic activity, while CelC and LicA are non-cellulomal proteins with β -1,4 and β -1,3 hydrolytic activities respectively. Thus, by combining various regulatory mechanisms it appears that *C. thermocellum* can regulate the expression of its carbohydrate active enzymes depending on available polysaccharides.



Putative GlyR3 binding site	SNP Sequence	Score	Strand
GTAAATCGGTTGCA	TGCAACCGATTTAC	6.07	-
	TGAAACCGATTTAC	10.60	-
	TGCAACCGATTTAA	10.25	-
	TGAAACCGATTTAA	15.51	-
TCAAAACAATTTCA	TCAAAACAATTTCA	7.09	+
	TGAAATCGTTTTGA	14.76	-

Figure 16 Putative binding site of *celT*

a) Intergenic region upstream of *celT*. Underlined sequence indicates putative GlyR3 binding sites identified by bioinformatics analysis. Arrow indicates apparent transcription start site identified by RNA-seq. b) Hypothetical SNPs that could substantially increase the strength of binding of the putative GlyR3 binding sites. Locations of SNP are underlined.

1.5 Methods

1.5.1 Bioinformatics analysis

The DBTBS transcriptional regulation database (http://dbtbs.hgc.jp) (Sierro et al., 2008) was used to identify 44 CcpA LacI binding sites in *B. subtilis* and a consensus sequence was determined. The frequency of each base at each of the 14 positions were normalized to determine the information content of the sequence (T. D. Schneider et al., 1986) and the position specific scoring matrix (PSSM) was determined. This matrix was used to search for possible GlyR3 binding sites in the proximity of the *manB-celT* gene cluster in *C. thermocellum*.

1.5.2 Bacterial strains

Bacterial strains and plasmids used in this study are summarized in Table 2. The *glyR3* gene was cloned into pTXB1 (New England Biolabs) expression vector and transformed into *E. coli* Top10 (Invitrogen). The plasmids were harvested and purified by Minipreparation kit (Wizard® Plus Minipreps, Promega) and transformed into T7 Express Competent *E. coli* strain C2566 (New England Biolabs) for production of GlyR3 (Table 2).

1.5.3 Culture conditions

C. thermocellum cultures were prepared under anaerobic conditions in 100 mL batch serum bottles and grown at 60°C in chemically defined MTC medium prepared as described by Zhang et al. (Zhang & Lynd, 2005). Avicel (PH105, FMC Biopolyer, Philadelphia, PA) was used as the carbon source. *E. coli* were grown in liquid culture with shaking at 37°C in Lunia-Bertani(LB) medium containing 100 µg/ml ampicillin. Expression of *glyR3* was induced with 0.5 mM isopropyl thiogalactoside (IPTG) when an OD_{600} of 0.4 was obtained. *E. coli* colonies for screening and selection were grown on LB medium agar with 100 µg/ml ampicillin at 37°C. *C. thermocellum* ATCC 27405 cultures used for RNA sequence analysis were grown in MTC medium in batch fermentations as described previously and for previously described samples (Yang et al., 2012).

Briefly, control cellobiose fermentations contained no ethanol supplementation while treatment fermentations were exposed to 3.9 g/L (or 0.5% [v/v]) ethanol shock at mid-exponential growth phase (OD_{600} ~ 0.5). Samples were taken pre-shock and 2, 12, 30, 60, 120, and 240 min post-shock from treated and untreated control fermentations.

Table 2 Strains and plasmids

Strains	Description	Reference or source
Clostridium thermocellum		
ATCC 27405	Wild type	ATCC 27405
Escherichia coli		
TOP10	F- mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 ΔlacX74 nupG recA1 araD139 Δ(ara-leu)7697 galE15 galK16 rpsL(Str^R) endA1 λ^-	(Newcomb et al., 2007a)
C2566	*fhuA2 lacZ::T7 gene1 [lon] ompT gal sulA11 R(mcr-73::miniTn10--*Tet^S*)2 [dcm] R(zgb-210::Tn10--*Tet^S*) endA1 Δ(mcrC-mrr)114::IS10*	This work
Plasmids		
pTXB1-glyR3	glyR3 cloned into pTXB1	(Newcomb et al., 2007a)
pCR2.1-TOPO		This work

1.5.4 Cloning of glyR3

Genomic DNA was extracted from *C. thermocellum* using the Wizard® Genomic DNA Purification Kit (Promega) and was used as a template for the amplification of the *glyR3* gene.The target DNA was PCR amplified using PuReTaq™ Ready-To-Go™ PCR beads (GE Health care) following reference (Newcomb et al., 2007a). *Eco*RV and *Xho*I were used for restriction sites (Newcomb et al., 2007a) at 37°C overnight after washing by MinElute PCR purification kit (Qiagen) (Table 3, primers 1 and 2). The target DNA was inserted into the pTBX1 plasmid (New England Biolabs) and transformed into *E. coli* (Oneshot Top10, Invitrogen). The colonies were selected on an ampicillin (100 µg/ml) plate. The target plasmid was extracted using a Minipreparation kit (Wizard® Plus Minipreps, Promega) and then sequenced for verification.

1.5.5 Expression and Purification of GlyR3

GlyR3 was obtained using an expression vector and purified following the procedures of reference (Newcomb et al., 2007a). The cloned vector was transformed into T7 Express Competent *E. coli* (Table 2) and induced with 0.5mM IPTG. Expression of the target protein was verified using SDS-PAGE. GlyR3 was purified following the IMPACT system protocol (New England Biolabs). The purified GlyR3 concentration was measured using the Bradford (Bio-rad) method with bovine serum albumin as a standard.

1.5.6 Electrophoretic mobility shift assay (EMSA)

DNA fragments from *celC*, *manB* and *celT* were amplified with biotin labeled primers 3-10 (Table 3). GlyR3 protein was obtained as described above. Running buffer was prepared following the LightShift Chemiluminescent EMSA kit Protocol. Electrophoresis was done under 100mV for 30 min with TBE (Tris-Bis-EDTA) gel (Invitrogen) which was transferred to nylon paper. The signal was developed using the LightShift Chemiluminescent EMSA kit. The image was detected by ChemiDOC XRS+(Bio-Rad) with 30 sec exposure. Unlabeled 18-mers matching the sequence of interest were used to competitively bind GlyR3 and provide confirmation of the binding locations. The effect of laminaribiose (Megazyme) on GlyR3-DNA interactions was also assessed using EMSA.

1.5.7 RNA extraction

To preserve RNA for qRT-PCR analysis, live cells were centrifuged at 10,000 g for 5 minutes. The cells were resuspended in10 volumes of RNAlater (Qiagen) and incubated for 15 minutes. The cells were centrifuged again and 1 mL of TRIzol (Invitrogen) was added before freezing at -80°C. After thawing, the cells were subjected to three 20-second cycles of bead beating (FastPrep-24, MP Biomedical). Chloroform (250 ml) was added to the lysate before vortexing for 45 seconds. The contents were centrifuged after a 3 minutes rest and the upper layer was collected and mixed with ethanol (1:1 volume). RNA was purified using an Ultra CleanTM Microbial RNA isolation kit (MO BIO). DNA

contamination was removed by addition of DNase I (Qiagen) to the membrane of the kit.

1.5.8 Quantitative Real Time PCR

Brilliant ®II SYBR Green QRT-PCR Master Mix kit (Agilent technologies) was used with 10 ng total RNA and 100 nM each forward and reverse primer as listed in Table 3 (17-22). New primers were designed using Primer3.

1.5.9 mRNA Sequencing

Transcriptomic profiles have been characterized previously using an expression DNA microarray (Yang et al., 2012). In this study, we generated RNA-Seq data from two independent biological replicates for each of the 60 min treatment and control samples using the 454 GS FLX instrument (454 Roche, Branford, CT) (Margulies et al., 2005). The 454 DNA libraries were generated from total RNA following the manufacturer's instructions, except that the small fragment removal step was omitted and the libraries were then sequenced using Titanium chemistry. Sequence data has been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession number SRX021497. Data deposited under SRR605557 and SRR605560 are derived from control fermentations 1 and 3, respectively, and data under SRR605564 and SRR605566 are derived from ethanol treatment fermentations 2 and 4, respectively.

1.5.10 In Vitro Transcription Assay

In this assay, the celC promoter and coding region and manB coding region which was generated by using primers 23-26 (Table 3) was inserted into pCR2.1-TOPO vector. The plasmid vector was transformed into TOP 10 *E. coli* to be selected on ampicillin contained LB plate. The DNA template was generated by using primers 23, 25 and 27 (Table 3) to amplify the promoter and coding region of *celC* and *manB*. TranscriptAid T7 High Yield Transcription Kit (Thermo scientific) was used to perform the *in vitro* transcription assay. DNase was treated followed by protocol from manufacturer. The RNA product was isolated by using the TRIzol method (Invitogen) along with RNeasy mini kit (Qiagen). Quantifying of mRNA expression was performed by 1-step SYBR Green QRT-PCR (Agilent Technologies) with the primers 17-20 (Table 3).

Table 3 Primers and probe

No	Sequence	Reference
1	F:glyR3-F-EcoRV-GCGCGATATCACCAGTGAAGAAATAGCAAAATTA	(Newcomb et al., 2007a)
2	R:glyR3-R-XhoI-GCGCCTCGAGGAATTCCAAAGCCCTCTTGGTT	(Newcomb et al., 2007a)
3	celCemsa-F-Biotin-CCGAATAAAAACTGGACAGAG	(Newcomb et al., 2007a)
4	celCemsa-R-Unlab TCCTCCTGAAATATTGTGTTTTA	(Newcomb et al., 2007a)
5	manBemsa-F-Biotin-TCGGTGAATGTTGAGGTTGA	This work
6	manBemsa-R-Unlab-TCCTGCTGAAATCTCTCTCCA	This work
7	celTemsa-F-Biotin-TGAGCAAATCAATTGTAATATGAAGA	This work
8	celTemsa-R-Unlab-TGCAGCAGAACGTTTCTTTC	This work
9	ControlF-Biotin-TGTTATCTTCGGTTAGCTCATCA	This work
10	ControlR-Unlab-TGGGAGGATGATACTGCTGTT	This work
11	celC-compt-F- AATGAACGCGCGTACATT	(Newcomb et al., 2007a)
12	celC-compt-R- AATGTACGCGCGTTCATT	(Newcomb et al., 2007a)
13	manB-compt-F- AATGTAAACGGTGTCAAT	This work
14	manB-compt-R- ATTGACACCGTTTACATT	This work
15	celT-compt-F- ATGTAAATCGGTTGCAGT	This work
16	celT-compt-F- ACTGCAACCGATTTACAT	This work
17	celCq-F-CGGGAACATATTGCCTTTGAAC	(Newcomb et al., 2007a)
18	celCq-R-GGTGGAATCAATTTCCCTGATTG	(Newcomb et al., 2007a)
19	Cthe_2811_F AACACATCCTTTCGGGTCAG	This work
20	Cthe_2811_R CACATCAAACCGTTCCCTCT	This work
21	Cthe_2812_F CTACAGCAGCAATTCGGTCA	This work
22	Chte_2812_R CATTGAAGAATCCCCGAAGA	This work
23	InvtCelC_F CCGAATAAAAACTGGACAGAAG	This work
24	InvtCelC_R CCAGTGGGCTTTCTGATGC	This work
25	InvtManB_F CGGCCAAATATGCCATAGAC	This work
26	InvtManB_R TCCTGAACCTGCTTGAGCTT	This work
27	InvtPCR_F CCATAAAACCGCCCAGTCTA	This work

CHAPTER II
GLYR3 AS A GLOBAL REGULATOR: REVEALS TWO MOTIFS FOR GLYR3 BOTH POSITIVE AND NEGATIVE REGULATION IN *CLOSTRIDIUM THERMOCELLUM*.

A version of this chapter will be submitted for publication by Jinlyung Choi and Chris Cox:

The manuscript will be submitted to Biotechnology for Biofuels.

Jinlyung Choi design and conduct experiment, analyze data and wrote paper.

2.1 Abstract

Clostridium thermocellum is an anaerobic, thermophilic, ethanogenic bacterium that has the ability to utilize cellulosic biomass. GlyR3, as a transcription factor, is known to regulate enzymes related to carbohydrate metabolism and transport and to autoregulate its own production through a laminaribiose-dependent mechanism in *Clostridium thermocellum*. We used the CcpA binding motif of *B. subtilis* to search for additional potential GlyR3 binding sites among *C. thermocellum* genes shown to change expression levels upon exposure to laminaribiose. Electrophoretic mobility shift assay (EMSA) and in vitro transcription (IVT) assays were performed on selected genes to provide additional evidence of a GlyR3-based mechanisms for these genes. We categorized genes showing strong evidence of regulation by GlyR3 into two groups according to whether they were up regulated or down regulated upon addition of laminaribiose. Each group of genes was characterized by a distinct binding motif: a *celC*-type motif associated with genes up regulated by laminaribiose and a *manB*-type motif associated with genes down regulated by laminaribiose, Together these lines of evidence suggest that, GlyR3 regulates up

to 32 genes in *Clostridium thermocellum* through laminaribiose-dependent mechanisms.

2.2 Introduction

Gram-positive, anaerobic, thermophilic, cellulolytic and ethanogenic abilities gives C. thermocellum advantages to playing an important role in ethanol production (Lynd et al., 2002a). The cellulosome is an enzyme system in *C. thermocellum,* which is principally responsible for its cellulolytic ability. The cellulosome has many proteins and has a different composition in different conditions, but its regulation needs more research. (Bayer et al., 2004a)

GlyR3 is a transcription factor known to regulate hydrolytic genes, such as *celC, licA* and *manB* (Newcomb et al., 2007a; Newcomb et al., 2011a). GlyR3 is a gene in the *celC* operon which is known to negatively autoregulate the operon (Newcomb et al., 2007a). The negative auto repression is relieved in the presence of laminaribiose. It is also known that the *celC* operon and *manB-celT* gene cluster are co-regulated (Newcomb et al., 2011a). The *celC* operon and *manB* are regulated in opposite directions with respect to the presence of laminaribiose: *celC* expression increases while *manB* expression decreases in comparison to conditions without laminaribiose. GlyR3 has been shown to be the factor controlling the regulation of both *celC* and *manB* through differences in binding affinities to the respective GlyR3 binding sites both in the presence and absence of laminaribiose (Choi et al.).

The protein binding domain and target DNA motif of GlyR3 are very similar to those of the global regulatory protein CcpA in *B. subtilis.* In this paper, we hypothesize that similar to CcpA, GlyR3 may also regulate multiple genes in *C. thermocellum* and that the CcpA binding motif may be used to identify putative binding sites for GlyR3. We selected 11 genes or operons that both changed regulation in the presence of laminaribiose and that contained a DNA sequence similar to the CcpA binding site. Through electrophoretic mobility shift assay (EMSA) we found that 7 of these DNA sequences bind laminaribiose. We performed *in vitro* transcription (IVT) assays on two of these genes to obtain additional evidence of direct regulation of these genes by GlyR3. We then classified these 7 as having either *celC*-type or *manB-type* responses to laminaribiose and identified distinct binding motifs for each type. Finally, we search for additional operons with strong *celC*-type and *manB-type* motifs in *C. thermocellum* are demonstrate that in the presence of laminaribiose, operons with *celC*-type binding sites are more likely to be upregulated, while operons with *manB*-type binding sites are more likely to be down regulated.

2.3 Results

2.3.1 Whole genome in vivo expression profiles by mRNA sequencing reveal that a large number of genes change expression in the presence of laminaribiose.

RNA sequencing was performed to measure the change in expression profile of the whole genome in the presence of laminaribiose. Triplicate samples were collected from experimental and control cultures 0, 1, 2 hours after adding laminaribiose and sent to DOE's Joint Genome Institute (JGI) for RNA sequencing as described in the methods. Differentially expressed genes at a significance level of $p<0.05$ were identified through ANOVA using the JMP genomics platform. Changes in expression were transformed to log base 2 and normalized using upper quartile scaling (UQS) prior to ANOVA analysis. Table 4 shows that 257 genes are induced and 402 genes are repressed at a significance level of $p<0.05$ (Additional File 4). Functional categories of amino acid transport and metabolism and translation, energy production and conversion, nucleotide transport and metabolism, ribosomal structure and biogenesis are induced and signal transduction mechanisms is repressed significantly as determined by having an odds-ratio greater than 2. The odds ratio compares the probability that a given outcome will occur under a particular condition to the probability that the outcome will occur in the absence of that condition (Bland & Altman, 2000).

Table 4 Gene category of change expression

Category	Total	Induced	Repressed	Odds ratio induced	Odds ratio repressed
Amino acid transport and metabolism	151	43	19	4.17	0.91
Carbohydrate transport and metabolism	114	14	17	1.47	1.11
Cellular Processes	456	26	89	0.63	1.54
Chromatin structure and dynamics, DNA replication, recombination and repair	160	10	23	0.70	1.06
Coenzyme metabolism	75	7	11	1.08	1.09
Energy production and conversion	93	20	11	2.87	0.85
General function prediction only	652	44	94	0.76	1.07
Hypothetical	741	33	70	0.49	0.66
Hypothetical/Unassigned	70	3	15	0.47	1.73
Lipid metabolism	30	2	4	0.75	0.98
Membrane Transport	2	0	0	0.00	0.00
Nucleotide transport and metabolism	56	12	5	2.86	0.62
Signal transduction mechanisms	93	4	23	0.47	2.08
Transcription, RNA processing and modification	108	9	12	0.95	0.79
Translation, ribosomal structure and biogenesis	150	30	9	2.62	0.40
Total*	2951	257	402		

*The differences in the listed total number of genes compared to sum values of each column are due to missing annotations for some genes.

2.3.2 The CcpA binding motif can be used to find additional GlyR3 binding sites

The position specific scoring matrix (PSSM) derived from the CcpA binding motif of *B. subtilis* was previously used to find a GlyR3 binding site within the manB coding region of *C. thermocellum* (Choi et al.). Here, we use the same PSSM to find additional putative GlyR3 binding sites associated with genes that are differentially expressed in the presence of laminaribiose. The receiver operating characteristic (ROC) curve for the PSSM for CcpA previously studied (Choi et al.) suggests that a PSSM score > 12 could be used in identifying new binding sites (sensitivity = 0.89 ; 1- specificity = 0.00069 ; true positive/false positive = 0.098). We searched for GlyR3 binding sites in the 400 bp upstream of the coding region or in the first 60% of the coding region of the first genes in operons. The operon structure of *C. thermocellum* is provided from DOOR2 (Mao et al., 2014). The chi-square test (SPSS, chi-square) was used to confirm that operons with PSSM scores greater than 12 were significantly enriched in genes changing expression in the presence of laminaribiose (sig = 0.013) compared to genes with scores below 12. The chi-square test results demonstrate that the PSSM for CcpA can be useful in finding GlyR3 binding sites even in a system likely characterized by a high false positive rate for GlyR3 binding and in a situation where genes are changing expression via mechanisms not related to GlyR3.

To reduce the false positive rate, we searched for GlyR3 binding sites only in operons characterized by a significant change in gene expression. A total of 34 genes showed both a significant change in expression and contained a potential GlyR3 binding site with a PSSM score of greater than 12 (Additional File 5). We choose 11 candidate genes that were characterized by a change in gene expression and with a variety of PSSM scores ranging from greater than 8 to greater than 14 for experimental investigation.

2.3.3 EMSA provides evidence of GlyR3 binding by putative binding sites.

Electrophoretic Mobility Shift Assay (EMSA) was performed to obtain evidence of direct binding between GlyR3 and the binding sites. EMSA experiments for confirmed binding sites in *celC* and *manB* were also included as positive controls and for comparison purposes to bring the total genes tested to 13. The results of all EMA experiments are summarized in Table 5 and gel blot images are provided in additional file 3. GlyR3 binding affinity is a ratio of the intensity of the shifted band over probe-only band as determined by ImageJ (C. A. Schneider, Rasband, & Eliceiri, 2012). Nine of the 13 genes show positive DNA-GlyR3 binding interactions, while the other four genes do not. The genes Cthe_2811, Cthe_0480, Cthe_2993, and Cthe_0720 shows similar affinity of binding to GlyR3 and all four genes show repressed gene expression. The genes Cthe_2807, Cthe_0391, Cthe_1332, Cthe_2714 and Cthe_3235 are all up regulated in the presence of laminaribiose and their bind sites show various

levels of positive interaction with GlyR3. Only Cthe_0391 shows strong binding even though it is weaker than Cthe_2807.

Additional EMSA experiments were conducted to assess the effect of laminaribiose on GlyR3 binding to identified binding sequences for Cthe_0391 and Cthe_0480. Cthe_0391 was observed to behave similarly to celC (Choi et al.; Newcomb et al., 2007a) in that GlyR3 did not bind to the target DNA sequences when laminaribiose was added. Cthe_0480 was observed to behave similarly to *manB* in that its GlyR3 binding behavior was not affected by addition of laminaribiose (Figure 17).

2.3.4 In vitro transcription assay reveals GlyR3 and laminaribiose directly impact expression level of Cthe_0391 and Cthe_0480

We used *in vitro* transcription assay to verify that the observed changes in gene expression were directly attributable to the interactions of GlyR3 with the identified binding sequences for Cthe_0391 and Che_0481. Messenger RNA was quantified by RT-PCR. RNA expression for Cthe_0391 was repressed for both dosages of GlyR3 ($p=0.01$ for 100 ng and $p=0.00003$ for 200 ng GlyR3). In the case of Cthe_0480 RNA expression was only repressed at the higher dosage of GlyR3 ($p=0.54$ for 100 ng and $p=0.015$ for 200 ng).

Table 5 Summary of EMSA experiments

Gene	PSSM Score (CcpA motif)	GlyR3 binding affinity (EMSA)	GlyR3 Binding response to laminaribiose (EMSA)	Expression
Repressed	**(manB-type)**			
Cthe_2811 (*manB*)[1]	14.62	0.37	No effect	1.49
Cthe_0480	14.47	0.24	No effect	0.78
Cthe_2993	8.48	0.38	Not tested	0.76
Cthe_0720	12.49	0.34	Not tested	0.42
Induced	**(celC-type)**			
Cthe_2807 (*celC*)[1]	9.19	0.69	Releases GlyR3	-1.41
Cthe_0391	12.69	0.46	Releases GlyR3	-0.55
Cthe_1332	12.90	0.30	Not tested	-0.31
Cthe_2714	13.45	0.17	Not tested	-0.34
Cthe_3235	13.68	0.31	Not tested	-0.54
No binding				
Cthe_3065	11.97	no binding	Not tested	0.93
Cthe_0298	14.47	no binding	Not tested	0.43
Cthe_1170	11.22	no binding	Not tested	0.42
Cthe_0422	9.20	no binding	Not tested	-1.23

[1](Choi et al.)

In Cthe_0391, the repression was released when laminaribiose was added (p=0.00059 for addition of laminaribiose to 200 ng GlyR3) but gene expression was not affected by laminaribiose addition in Cthe_0480 (p=0.16 for addition of laminaribiose to 200 ng GlyR3) (Figure 18). A negative control experiment (shown in additional file 9) demonstrated that the addition of laminaribiose in the absence of GlyR3 had no effect on gene expression. The behavior of both genes was consistent with their respective GlyR3-DNA binding behavior observed during EMSA and with their respective *in vivo* gene expression patterns in the presence of laminaribiose. The overall behavior of Cthe_0391 and Cthe_0480 are similar to *celC* and *manB*, respectively.

2.3.5 Up and down regulated genes have distinct binding motifs.

The similarity of gene regulation behavior between celC (Cthe_2807) and Cthe_0391 suggest that they are regulated by a common mechanism in which laminaribiose interferes with the binding between GlyR3 and its DNA binding site such that expression is increased in the presence of laminaribiose (Choi et al.; Newcomb et al., 2007a). In addition, we hypothesize that other upregulated genes (Cthe_1332, Cthe_2714, Cthe_3235 and Cthe_2807) may also be regulated by a similar mechanism and tentatively classify these as celC-type binding sites (Table 5). Conversely, *manB* (Cthe_2811) and Cthe_0480 genes are down regulated in the presence of laminaribiose (Choi et al.; Newcomb et al., 2007a). In this case, the EMSA, and *in vivo* and *in vitro* expression experiments

are consistent with a mechanism in which the binding sites may only bind GlyR3 when the concentration is relatively high. In addition, the GlyR3-DNA interactions may be relatively unaffected by direct interaction with laminaribiose (Choi et al.). We hypothesize that other down-regulated genes (Cthe_2993, and Chte_0720) may also be regulated by similar mechanisms and tentatively classify these as *manB*-type binding sites (Table 5).

We hypothesized that the differences in gene expression behavior and regulatory mechanisms for *celC*-type and *manB*-type genes as summarized in Table 5 might be attributable to subtle differences in their binding motifs. We generated separate PSSMs and sequence logos for the 5 *celC*-type and 4 *manB*-type genes in Figure 3B and 3C. Both *celC*-type and *manB*-type motifs are generally similar to each other and to the CcpA motif in *B. subtilis* (Figure 3A). Overall, the *celC*-type motif is more similar to the CcpA motif, while the *manB*-type motif shows more variability.

CelC is characterized by strongly conserved GCGC sequence at positions 6-9 while the *manB*-type motif has a consensus of ACGC in the same positions with only the CG in positions 7 and 8 being strongly conserved. The *celC*-type motif is also characterized by strongly conserved bases of TG at positions 1 and 2 and a TCA sequence at positions 12-14. In comparison, the *manB*-type motif has same consensus sequences in these positions, but allows for much more variability in these positions.

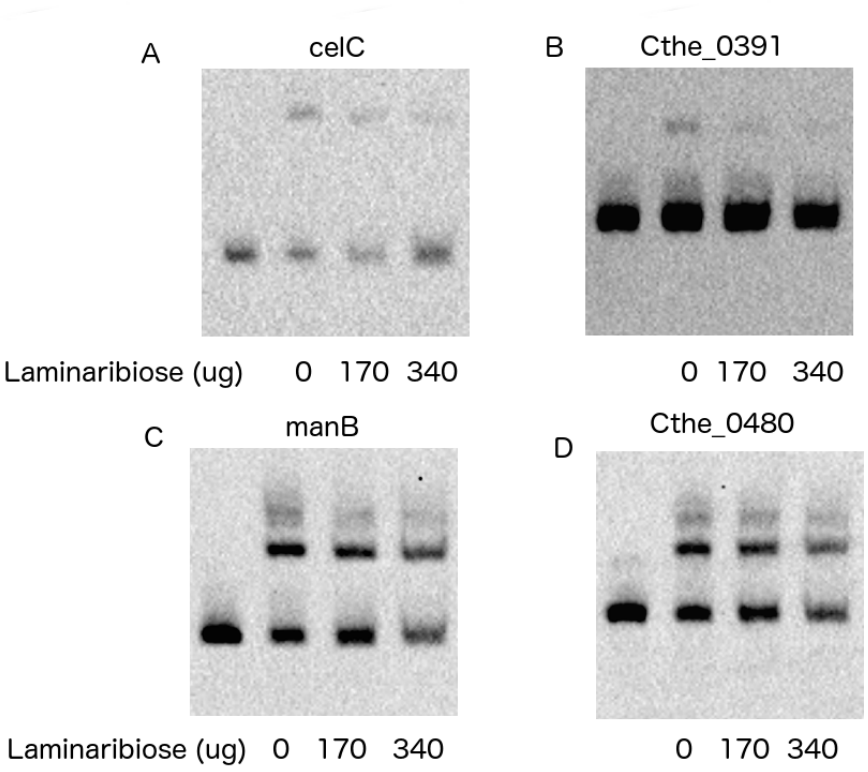


Figure 17 Laminaribiose effect of DNA-GlyR3 binding revealed by EMSA;

Electrophoretic Mobility Shift Assay shows binding of GlyR3 to Cthe_0480 and Cthe_0391. All lanes contain 0.2 ng of DNA (celC: 0.3 nM, Cthe_0391: 0.2 nM, manB 0.18 nM, Cthe_0480: 0.13 nM) and 20 ng of GlyR3 (51.59 nM). Laminaribiose: 170 ug (20 mM), 340 ug (50 mM)

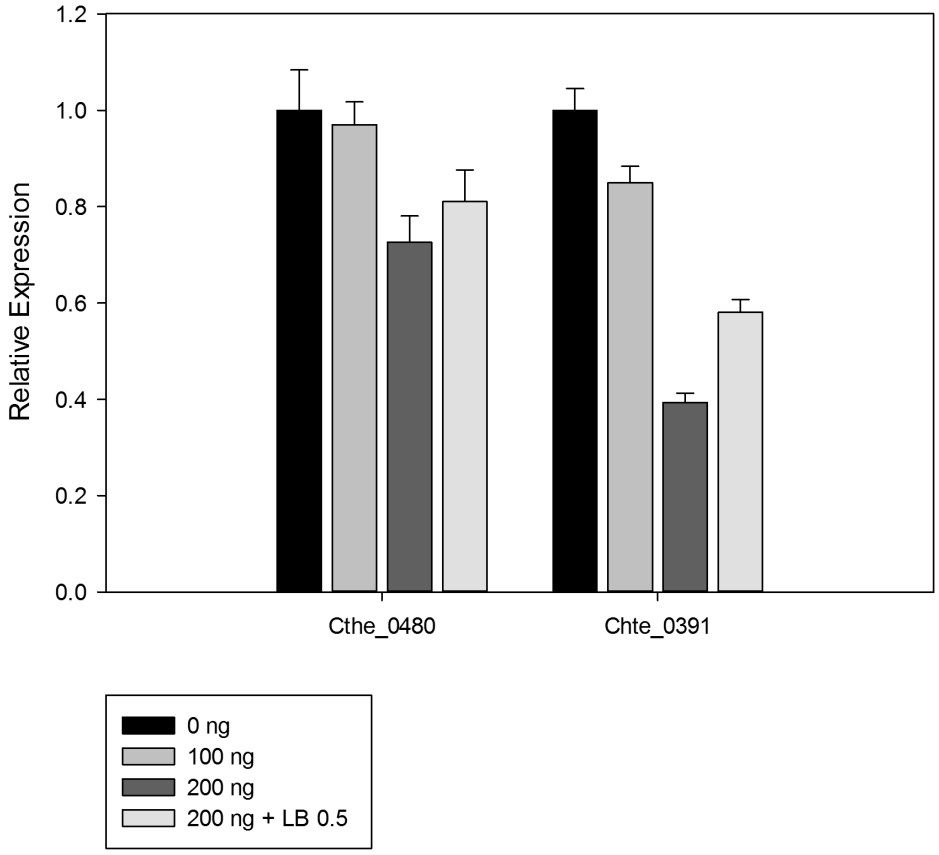


Figure 18 *in vitro* transcription assay

Relative gene expression of Cthe_0480 and Cthe_0391 with and without adding 100 ng of GlyR3 (257.9 nM), 200 ng (515.9 nM) of GlyR3, and 200 ng of GlyR3 and 0.5 M (50 mM final concentration) of laminaribiose determined by qRT-PCR. DNA concentration: Cthe_0480 (0.7 nM), Cthe_0391 (0.8 nM)

Despite these differences, the *celC*-type and *manB*-type motifs are similar, and the question arises whether these two motifs are significantly different. (Figure 19B and C) To answer this question, we randomly generated 100 14 base-pair sequences from the PSSM of each motif. Each random sequence is scored by *celC*-type motif and *manB*-type motif. If the score from *celC*-type PSSM is higher than score from *manB*-type PSSM, the sequence is assigned as *celC*-type, and vice versa. The sequences that are assigned correctly are counted. Whole process was repeated three times. On average 92 artificial sequences (standard deviation = 2) are assigned correctly to the *celC*-type motif and 94.7 artificial sequences (standard deviation = 2.5) are assigned correctly to the *manB*-type motif. On this basis, we concluded that the motifs were significantly different and capable of being distinguished one from the other through their PSSM.

Next, we hypothesized that the genes having strong *celC*-type sites should be enriched in up-regulated genes, while those having strong *manB*-type binding sites should be enriched in down-regulated genes. All intergenic regions upstream of operons (400 bp of upstream if former gene is located more than 400 bp before) and the first 60% of the coding region of the first gene in the operon were searched with *celC*-type and *manB*-type PSSMs. The maximum score for a given operon and motif over all positions was chosen. The rates of induced expression of each increment of PSSM score were compared with rates of repression (Figure 19D). Operons that had *celC*-type PSSM scores greater

than 14 were enriched in up-regulated genes at a significance level of $p<0.006$ using Monte Carlo simulation of Fisher's exact test (SPSS, chi-square). Six operons are found to have PSSM scores above 14 and significant increases in gene expression. Four of the five *celC*-type operons listed in Table 5 were found using this method (Cthe_2714 was not found), in addition to two newly identified operons (Cthe_1312 and Chte_2619) were found. Cthe_1312 is a glycyl-tRNA synthetase and Cthe_2619 is a cell shape determining protein (MreB/Mrl family). For manB-type binding sites, the rates of induced and repressed operons at each score was also compared (Figure 19E). Operons that had *manB*-type PSSM scores greater than 12 were enriched in up-regulated genes at a significance level of $p<0.025$ using chi-square test (SPSS, chi-square). Twenty-six operons are found to have manB-type PSSM scores greater than 12 and significantly repressed gene expression. Three of the four operons in Table 5 were identified (Cthe_2993 was not) using this method. In addition 23 new operons were identified as listed in additional file 10. The functions of these genes will be discussed below.

2.4 Discussion

Motif-based searches have been used by many researchers to search for many genetic regulatory features, such as transcription factor binding sites, signaling pathways, RNA polymerase sigma factor binding sites, histone binding sites in eukaryotes, and other kinds of pattern recognition in DNA sequences (Bailey et al., 2009; Fouts et al., 2002; Lesburg et al., 1999; Sandelin, Alkema,

Engstrom, Wasserman, & Lenhard, 2004; Shen-Orr, Milo, Mangan, & Alon, 2002; Strahl & Allis, 2000; Yaffe et al., 2001).

Yet, motifs are only known for a relatively small number of transcription and sigma factors in a few organisms. To generate motifs, several experimentally determined binding sites must be know. EMSA and DNA footprinting are possible way to get information experimentally, but require significant time and effort.

Some of this effort can be avoided by using the known sequence information of related motifs in other organisms. CcpA is a one of the most well-known transcription factors in Gram-positive bacteria, which regulates carbon metabolism. CcpA and GlyR3 are both lacI family proteins that have helix-turn-helix domains. These two proteins have DNA binding domains with very high similarity in their amino acid sequences. An interesting characteristic of GlyR3 is its apparent ability to up regulate certain genes with one binding motif while down regulating genes with a different motif in response to the concentration of laminaribiose.

The regulatory action of GlyR3 can be compared to Leucine responsive protein (Lrp), a global regulator protein that regulates biosynthesis of leucine and other amino acids (Calvo & Matthews, 1994; Fraser & Newman, 1975). Lrp protein is known to have four different binding motifs and is known to be both a positive and a negative regulator. In salmonella, *fimA* and *fimZ* genes are induced only when Lrp is present at low concentration.

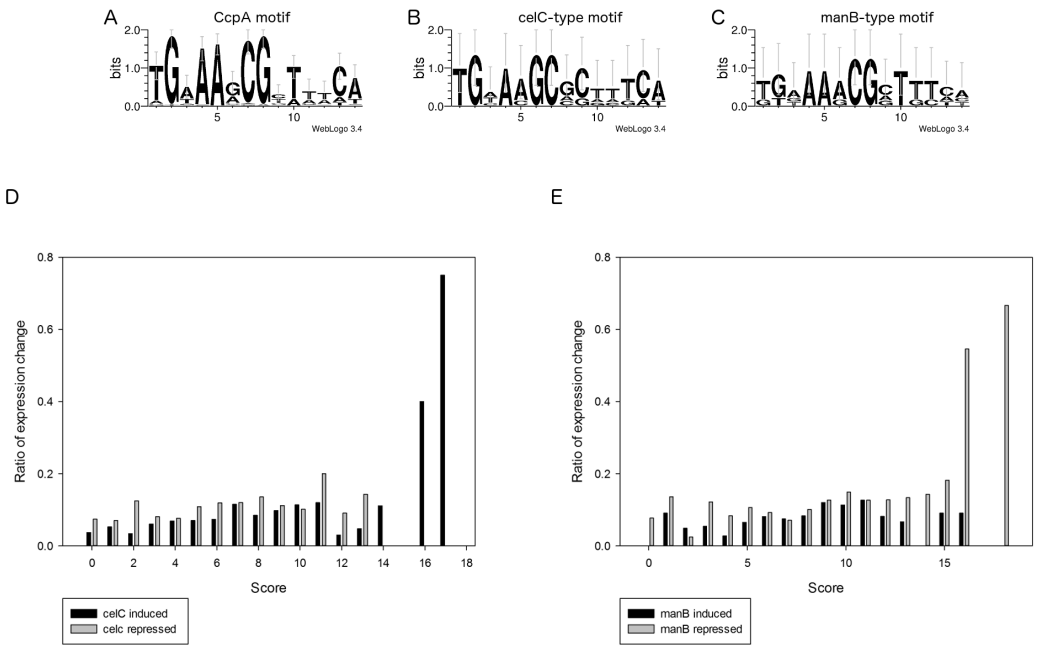


Figure 19 Logos of motifs and model validation

(a) CcpA motif, (b) *celC*-type motif, (c) *manB*-type motif are generated from Weblogo (Crooks et al., 2004). (d) Rate of change expression at each bins of scores from *celC*-type motif (e) and *manB*-type motif.

Those genes are repressed when Lrp is present at high concentration or when Lrp is not present(Baek, Kang, Roland, & Curtiss, 2011; Baek, Wang, Roland, & Curtiss, 2009). CcpA, a global regulator in *B. subtilis* is also known to be both a positive and negative regulator. For example, CcpA activates *ackA* expression and represses *phoP* expression (Grundy, Waters, Allen, & Henkin, 1993; Puri-Taneja, Paul, Chen, & Hulett, 2006).

We found that GlyR3 regulates two groups of genes with different regulation patterns. One group, called *celC*-type, has induced gene expression under presence of laminaribiose. Also, two of the genes in the group show inhibition of binding affinity in the presence of laminaribiose. Another group, called *manB*-type, has repressed gene expression under presence of laminaribiose. Two of genes in the group show that the binding affinity is not affected by presence of the laminaribiose.

The repressed genes with score over 12 with *manB*-type motif have function as 1 amino acid transport and metabolism, 2 carbohydrate transport and metabolism, 5 cellular processers, 3 chromatin structure and dynamics, DNA replication, recombination and repair, 2 energy production, 5 general function, 2 signal transduction, 6 hypothetical. The hypothetical protein (Cthe_0480) is a leading protein of long operon in which most of other genes have flagella synthesis function. So, it is a reasonable guess that the regulation of Cthe_0480 affects the expression of the downstream flagella genes (Figure 20). It is possible that the cell represses flagella synthesis when it finds new carbon source.

The activated genes with score 15 with *celC*-type motif have function as 2-carbohydrate transport and metabolism, 1 translation ribosomal structure and biogenesis, 1 hypothetical, 1 cellular processes, 1 chromatin structure and dynamics DNA replication recombination and repair.

The results of this research show that GlyR3 has ability to regulate genes in two different ways. One group of genes is induced and another group of genes are repressed under conditions of high GlyR3 expression. However, not all genes that changed expression upon addition of laminaribiose are regulated by GlyR3. It is possible that other unknown factors can be involved in the regulon. These unknown factors can be found by high throughput binding experiment such as ChiP-Seq data(Valouev et al., 2008).

2.5 Conclusions

Motif based search is a useful tool to search for binding sites in the whole genome. But experimental data is required to build motif.

We use the known motif of CcpA that is highly similarity with GlyR3 binding domain as a starting point. EMSA and IVT were performed to confirm the binding sites. Finally we can build up two motifs for GlyR3 binding.

Table 6 Odds ratio of celC-type and manB-type operon

Category	Total	celC-type	manB-type	celC-type odd ratio	manB-type odd ratio
Amino acid transport and metabolism	151	0	1	0.00	0.75
Carbohydrate transport and metabolism	114	2	2	8.76	2.01
Cellular Processes	456	1	5	1.08	1.25
Chromatin structure and dynamics, DNA replication, recombination and repair	160	0	3	0.00	2.15
Coenzyme metabolism	75	0	0	0.00	0.00
Energy production and conversion	93	0	2	0.00	2.47
General function prediction only	652	0	5	0.00	0.87
Hypothetical	741	1	5	0.66	0.76
Hypothetical/Unassigned	70	0	1	0.00	1.63
Lipid metabolism	30	0	0	0.00	0.00
Membrane Transport	2	0	0	0.00	0.00
Nucleotide transport and metabolism	56	0	0	0.00	0.00
Signal transduction mechanisms	93	0	2	0.00	2.47
Transcription, RNA processing and modification	108	0	0	0.00	0.00
Translation, ribosomal structure and biogenesis	150	2	0	6.63	0.00
Total	2951	6	26		

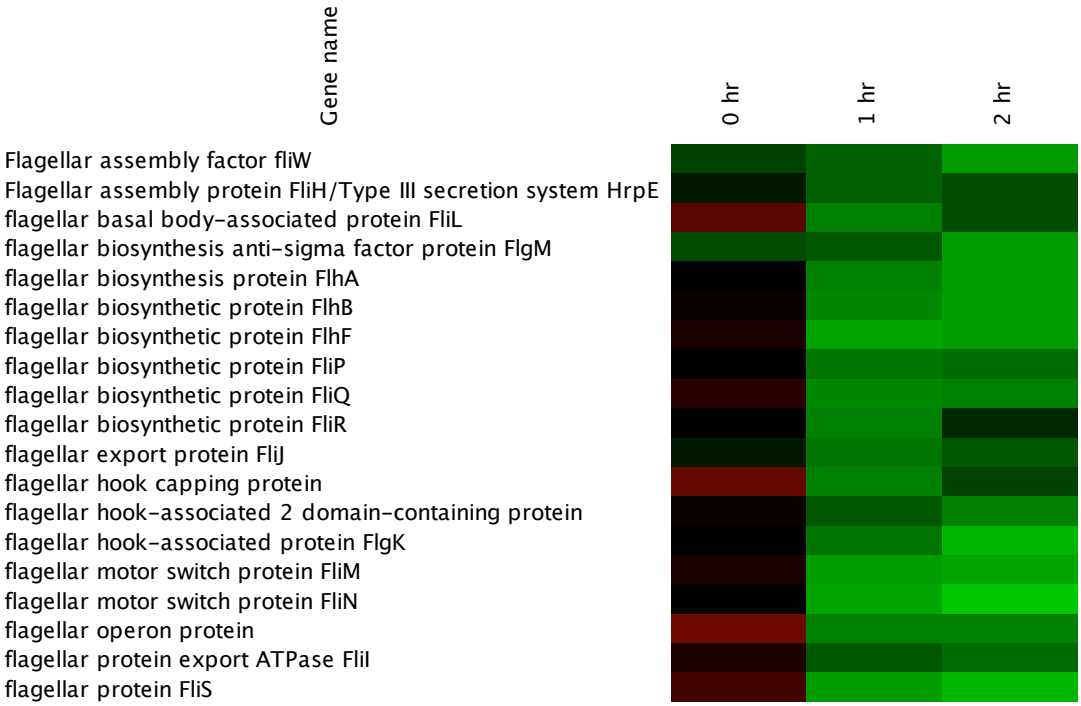


Figure 20 Heat map of flagella related genes

2.6 Methods

2.6.1 Bacterial strains and Culture Conditions

Bacterial strains and plasmids used in this study were described previously (Choi et al.) except that Zymo 5α (Zymo research) was used for this research(Additional File 8). *C. thermocellum* was grown at 60 °C in 100 mL batch serum bottles under anaerobic conditions in MTC medium(Zhang & Lynd, 2005) with Avicel (PH105, FMC Biopolymer, Philadelphia, PA) added as the sole carbon source. Cells were grown until late exponential phase before the addition of 1 mM laminaribiose. Triplicate cell samples were collected from control (no laminaribiose) and experimental (laminaribiose added) serum bottles at 0, 1 and 2 hours after the addition of laminaribiose.

2.6.2 RNA extraction

RNA extraction was performed according to the modified protocol as described previously(Choi et al.). Cell cultures 100 ml in volume were spun down at 8000g for 10 min at 4^{o}C in two 50 ml centrifuge tubes. The supernatant was discarded and 3 ml of Trizol was added to each pellet and vortexed vigorously. The cell pellet with Trizol was added into six 2 ml tubes with 0.6 ml of beads (Zirconia/Silica Beads, 0.1 mm dia, Biospec Products Inc.). The cells in other four tubes were beaten for 20 sec three times. The lysate was transferred into a new tube without beads. Chloroform (250 ul) was added, vortexed 45 sec, and incubated at room temperature for 3 min. The tubes were centrifuged at 10,000g

for 15 min at 4^{o}C. Upper phase (approximately 450 ul) was transferred into new 1.5 ml micro centrifuge tubes with 80% ethanol added with 1:1 ratio (v:v). The ethanol-sample mixtures in six tubes were loaded into same filter tube (repeat four times of loading and centrifuge). The following procedures were done according to the previous paper (Choi et al.). Total RNA samples were quality-controlled (Clear peak of 16s 23s and no sign of degradation) by bioanalyzer (Agilent Technologies). The RNA concentration was measured by Nano-drop (Thermo NanoDrop ND-1000). A total 18 RNA samples were sent to DOE's Joint Genome Institute for RNA sequencing.

2.6.3 RNA sequencing experiments

rRNA was removed from 1µg of total RNA using Ribo-Zero™ rRNA Removal Kit (Bacteria) (Epicentre). Libraries were generated using the Truseq Stranded mRNA Sample Preparation Kit (Illumina). Briefly, the rRNA-depleted RNA was fragmented and reversed transcribed using Superscript II (Invitrogen), followed by second strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing, adapter ligation and 10 cycles of PCR amplification. Libraries were quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then multiplexed into pools of 4 libraries each, and the pools were prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v3, and Illumina's cBot instrument to

generate clustered flowcells for sequencing. Sequencing of the flowcells was performed on the Illumina HiSeq2000 sequencer using a TruSeq SBS sequencing kit 200 cycles, v3, following a 2x100 indexed run recipe

2.6.4 RNA sequence analysis

RNA sequence analysis was performed to see the gene expression change for whole genome. RNA expression was counted by JGI. RNA expression counts are analyzed by JMP Genomics 10. The counts are transformed by log base 2, normalized by Upper Quartile Scaling(Linville et al., 2013). Differential expression was test statistically by an analysis of variance (ANOVA) considering time and concentration of laminaribiose to be the independent variables. A gene was considered statistically significantly differentially expressed if p value is less than 0.05.

2.6.5 Real-time qRT-PCR analysis

To further validate RNA-sequencing data, 11 genes were analyzed using real-time quantitative RT-PCR. Eleven genes are representing different range of gene expression based on RNA-sequencing data and primer pairs were designed using the program Primer 3 listed in additional file(Untergasser et al., 2012). Manufacture's standard protocol was followed (Brilliant II SYBR® Green QRT-PCR AffinityScript 2-Step Master Mix). The 1 hour with laminaribiose / 1 hour without laminaribiose ratio was calculated based on comparative ct method.

The linear correlation between the RT-PCR and RNA sequencing was calculated based on the log base 2 values. Gene expression measurement using RT-PCR and RNA sequencing were highly correlated with an R^2 value of 0.92 (Additional File 1)

2.6.6 Motif Based GlyR3 binding search

CcpA binding motif was used to search GlyR3 binding site. Similarity of CcpA and GlyR3, and calculating position specific scoring matrix (PSSM) is described previously(Choi et al.). Whole genome of *C. thermocellum* was searched with PSSM. Also, reverse complement sequences were compared with forward sequence.

The search results were merged into gene expression data from RNA sequencing after filtered by score (>12). The highest scored binding candidate is chosen in the location of upstream 400 bp and 60% of the early coding region on the first gene in the operon.

The whole operon search is repeated after generating celC-type and manB-type motif. The PSSM of two motifs are calculated using same method described previously. The list of operon is provided from DOOR2(Mao et al., 2014). The highest scored sequence is chosen among upstream of operon and 60% of the early coding region on the first gene in the operon for each motif. The search result is merged by gene expression change.

2.6.7 Electrophoretic mobility shift assay (EMSA)

EMSA were performed using LightShift Chemiluminescent EMSA Kit as described before(Choi et al.) except primers 23-48 (Additional File 2).

2.6.8 In vitro transcription assay (IVT)

Coding region of Cthe_0391 and Coding region of Cthe_0481 and putative GlyR3 binding site (last part of coding region of Cthe_0480) was amplified by PCR (illustra PuReTaq Ready-To-Go PCR Beads, GE healthcare) in thermocycler (S1000™ Thermal Cycler, bio-rad) with primer 49-52 (Additional File 2). PCR product was inserted into pCR2.1-TOPO vector using TOPO® TA cloning kit (Invitrogen). The cloned vector was transformed into E. coli strain Zymo 5α competent cell (Zymo research). The cell was plated on the LB media contain ampicillin and X-gal for blue-white screening. Few colonies are transferred into 50 ml LB liquid media for further selection then plasmid was harvested using Zyppy™ Plasmid Miniprep Kit (Zymo research). DNA fragment was amplified with forward primer 57 (Additional File 2) and each reverse primer 50 and 52 (Additional File 2) using PCR to select reverse insertion. Reverse insertion dose not show band when amplified DNA fragment run on the gel. Template DNA for in-vitro transcription assay was amplified with forward primer 58 and reverse primer 50 and 52 (Additional File 2). PCR product was washed using MinElute-PCR-Purification-Kit (Qiagen). In-vitro transcription assay was performed using TranscriptAid T7 High Yield Transcription Kit. RNA was

extracted using Trizol method described above. Quantification was performed using Brilliant II SYBR® Green QRT-PCR AffinityScript 1-Step Master Mix with primer 53-56 (Additional File 2).

2.6.9 Statistical validation

Receiver operating characteristic (ROC) curve was used for validating *ccpA* motif(Erill & O'Neill, 2009). Histogram of score of 44 known binding site and total operon(Mao et al., 2014) in *B. subtilis* is counted for each score. ROC curve is plotted followed the method by Erill and O'Neill(Erill & O'Neill, 2009). The score of a 1/10 true to false positive ratio is chosen for further search.

The Chi-square test (SPSS, chi-square) was used for validating CcpA motif. The score over 12 is considered as CcpA-site. The Fisher's exact test (SPSS, chi-square) with Monte Carlo was used for validating celC-type and manB-type motifs.

CONCLUSION

In this research, we studied GlyR3 regulation in *C. thermocellum*. Understanding of gene-regulation is an essential step to optimize ethanol production. GlyR3 is a transcription factor that regulates celC operon.

In chapter 1, we identified an additional GlyR3 binding site within the coding region of *manB*. This putative identification is based on the information content of CcpA binding site in *B. subtilis* because CcpA and GlyR3 have high structural similarity, *C. thermocellum* and *B. subtilis* are both gram-positive and both proteins regulate genes associated with carbon metabolism. The ability of GlyR3 to physically bind to the identified sequence was demonstrated used EMSA. An *in vitro* transcription (IVT) assay showed that GlyR3 was able to decrease the expression of *manB* without other factors. Both the EMSA and IVT assay results for *manB* were unaffected by addition of laminaribiose, which is contradictory to the affect of laminaribiose on the *celC* operon. This suggests that the interactions of sugar and the protein are dependent on the DNA binding sequence.

The *in vivo* gene expression profile determined by RT-PCR shows that celC operon and *manB* gene is regulated in opposite directions in the presence of laminaribiose. The celC operon is activated after adding laminaribiose and its induction is proportional to concentration of laminaribiose. However, *manB* is repressed under presence of laminaribiose but its change in expression level is not affected by different concentration of laminaribiose. EMSA and IVT results show that those expression changes are consistent with the *in-vivo* results and

explain how GlyR3 regulates two genes in different ways. At low concentrations of GlyR3, which occur without presence of laminaribiose, the *celC* operon is in the off-status because small amounts of GlyR3 are sufficient to negatively autoregulate the *celC* operon. When laminaribiose is added, the *celC* operon is activated because laminaribiose inhibits GlyR3 binding. High expression of GlyR3 can repress *manB* since the binding of GlyR3 to *manB* is not affected by laminaribiose.

In chapter 2, we hypothesize that GlyR3 may regulate additional operons in *C. thermocellum* other than *celC* operon and *manB*. The results in chapter 1 show that it is possible that GlyR3 may have the ability to regulate more than one gene. Genome wide transcriptomic analysis shows the extent of the gene expression change with the addition of laminaribiose, which causes over-expression of GlyR3. Messenger-RNA sequencing was performed with and without adding laminaribiose since the expression of GlyR3 is induced under the presence of laminaribiose. A selected number of putative GlyR3 binding sites that had high PSSM scores and showed significant change in gene expression were selected for addition study using EMSA and IVT to provide further evidence that these genes were regulated by GlyR3. These experiments suggested that the direction of regulation depends on the genes behavior with respect to the presence of GlyR3 and laminaribiose. Genes that were up regulated behaved similarly to the celC operon while genes that were down regulated behaved similarly to the manB operon. We then hypothesized that the differences in

behavior were directly related to differences in the GlyR3 binding sequences of the two groups of genes. Two new motifs, a celC-type motif associated with up regulated genes and a manB-type motif associated with down regulated genes were revealed. The celC-type motif was found to be significantly associated with genes up-regulated in the presence of laminaribiose, while the manB-type motive was found to be significantly associated with genes that were down regulated by laminaribiose.

A whole genome search of *C. thermocellum* using the celC-type motif and manB-type motif show that carbohydrate transport and metabolism related genes are possibly regulated by GlyR3 in both motifs more than any other category of function. There is an unexpected finding that GlyR3 may regulate genes involved in flagella synthesis.

However, it is difficult to say that all genes that have changed expression in the transcriptomic data were regulated only by GlyR3. It is likely that other unknown factors were affected by laminaribiose. Also, genes identified by binding motif search include false positives.

BIBLIOGRAPHY

Argyros, D. A., Tripathi, S. A., Barrett, T. F., Rogers, S. R., Feinberg, L. F., Olson, D. G., . . . Caiazza, N. C. (2011). High Ethanol Titers from Cellulose by Using Metabolically Engineered Thermophilic, Anaerobic Microbes. *Appl Environ Microbiol, 77*(23), 8288-8294. doi: 10.1128/aem.00646-11

Atsumi, S., Wu, T. Y., Machado, I. M. P., Huang, W. C., Chen, P. Y., Pellegrini, M., & Liao, J. C. (2010). Evolution, genomic analysis, and reconstruction of isobutanol tolerance in Escherichia coli. *Molecular Systems Biology, 6*. doi: Artn 449

Doi 10.1038/Msb.2010.98

Baek, C. H., Kang, H. Y., Roland, K. L., & Curtiss, R. (2011). Lrp Acts as Both a Positive and Negative Regulator for Type 1 Fimbriae Production in Salmonella enterica Serovar Typhimurium. *Plos One, 6*(10). doi: ARTN e26896

DOI 10.1371/journal.pone.0026896

Baek, C. H., Wang, S. F., Roland, K. L., & Curtiss, R. (2009). Leucine-Responsive Regulatory Protein (Lrp) Acts as a Virulence Repressor in Salmonella enterica Serovar Typhimurium. *Journal of bacteriology, 191*(4), 1278-1292. doi: Doi 10.1128/Jb.01142-08

Bahari, L., Gilad, Y., Borovok, I., Kahel-Raifer, H., Dassa, B., Nataf, Y., . . . Bayer, E. A. (2011a). Glycoside hydrolases as components of putative carbohydrate biosensor proteins in *Clostridium thermocellum*. *J Ind Microbiol Biotechnol, 38*(7), 825-832. doi: Doi 10.1007/S10295-010-0848-9

Bahari, L., Gilad, Y., Borovok, I., Kahel-Raifer, H., Dassa, B., Nataf, Y., . . . Bayer, E. A. (2011b). Glycoside hydrolases as components of putative carbohydrate biosensor proteins in *Clostridium thermocellum*. *Journal of Industrial Microbiology & Biotechnology, 38*(7), 825-832. doi: Doi 10.1007/S10295-010-0848-9

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., . . . Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research, 37*, W202-W208. doi: Doi 10.1093/Nar/Gkp335

Bayer, E. A., Belaich, J. P., Shoham, Y., & Lamed, R. (2004a). The cellulosomes: Multienzyme machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol, 58*, 521-554. doi: 10.1146/annurev.micro.57.030502.091022

Bayer, E. A., Belaich, J. P., Shoham, Y., & Lamed, R. (2004b). The cellulosomes: Multienzyme machines for degradation of plant cell wall polysaccharides. *Annual Review of Microbiology, 58*, 521-554. doi: 10.1146/annurev.micro.57.030502.091022

Bhandiwad, A., Shaw, A. J., Guss, A., Guseva, A., Bahl, H., & Lynd, L. R. (2013). Metabolic engineering of Thermoanaerobacterium saccharolyticum for n-

butanol production. *Metab Eng, 21C*, 17-25. doi: 10.1016/j.ymben.2013.10.012
Bland, J. M., & Altman, D. G. (2000). Statistics notes - The odds ratio. *British Medical Journal, 320*(7247), 1468-1468. doi: Doi 10.1136/Bmj.320.7247.1468
Brown, S. D., Lamed, R., Morag, E., Borovok, I., Shoham, Y., Klingeman, D. M., . . . Bayer, E. A. (2012). Draft Genome Sequences for *Clostridium thermocellum* Wild-Type Strain YS and Derived Cellulose Adhesion-Defective Mutant Strain AD2. *J Bacteriol, 194*(12), 3290-3291. doi: 10.1128/jb.00473-12
Calvo, J. M., & Matthews, R. G. (1994). The Leucine-Responsive Regulatory Protein, a Global Regulator of Metabolism in Escherichia-Coli. *Microbiological Reviews, 58*(3), 466-490.
Carere, C. R., Sparling, R., Cicek, N., & Levin, D. B. (2008). Third generation biofuels via direct cellulose fermentation. *International J. Mol Sci, 9*(7), 1342-1360. doi: 10.3390/ijms9071342
Ceres, I. (2014). Biofuels - Carbohydrates.
Choi, J., Klingeman, D. M., Brown, S. D., & Cox, C. D. The LacI family protein GlyR3 co-regulations the celC operon and manB in Clostridium thermocellum. *To be submitted*.
Chundawat, S. P. S., Beckham, G. T., Himmel, M. E., & Dale, B. E. (2011). Deconstruction of Lignocellulosic Biomass to Fuels and Chemicals. *Annual Review of Chemical and Biomolecular Engineering, 2*(1), 121-145. doi: doi:10.1146/annurev-chembioeng-061010-114205
Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research, 14*(6), 1188-1190. doi: Doi 10.1101/Gr.849004
David, K., & Ragauskas, A. J. (2010). Switchgrass as an energy crop for biofuel production: A review of its ligno-cellulosic chemical properties. *Energy & Environmental Science, 3*(9), 1182-1190.
Demain, A. L., Newcomb, M., & Wu, J. H. D. (2005). Cellulase, clostridia, and ethanol. *Microbiol Mol Biol Rev, 69*(1), 124-+. doi: 10.1128/mmbr.69.1.124-154.2005
Erill, I., & O'Neill, M. (2009). A reexamination of information theory-based methods for DNA-binding site identification. *BMC Bioinformatics, 10*(1), 57.
Fontes, C., & Gilbert, H. J. (2010). Cellulosomes: Highly Efficient Nanomachines Designed to Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. In R. D. Kornberg, C. R. H. Raetz, J. E. Rothman & J. W. Thorner (Eds.), *Annual Review of Biochemistry, Vol 79* (Vol. 79, pp. 655-681).
Fouts, D. E., Abramovitch, R. B., Alfano, J. R., Baldo, A. M., Buell, C. R., Cartinhour, S., . . . Collmer, A. (2002). Genomewide identification of Pseudomonas syringae pv. tomato DC3000 promoters controlled by the

HrpL alternative sigma factor. *Proceedings of the National Academy of Sciences of the United States of America, 99*(4), 2275-2280. doi: Doi 10.1073/Pnas.032514099

Fraser, J., & Newman, E. B. (1975). Derivation of Glycine from Threonine in Escherichia-Coli K-12 Mutants. *Journal of bacteriology, 122*(3), 810-817.

Fried, M., & Crothers, D. M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Research, 9*(23), 6505-6525. doi: 10.1093/nar/9.23.6505

Garciamartinez, D. V., Shinmyo, A., Madia, A., & Demain, A. L. (1980). Studies on Cellulase Production by *Clostridium Thermocellum*. *Eur J Appl Microbiol d Biotechnol, 9*(3), 189-197. doi: 10.1007/bf00504485

Giampietro, M., Ulgiati, S., & Pimentel, D. (1997). Feasibility of large-scale biofuel production - Does an enlargement of scale change the picture? *Bioscience, 47*(9), 587-600. doi: Doi 10.2307/1313165

Gilbert, H. J. (2007). Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. *Mol Microbiol, 63*(6), 1568-1576. doi: 10.1111/j.1365-2958.2007.05640.x

Grundy, F. J., Waters, D. A., Allen, S. H. G., & Henkin, T. M. (1993). Regulation of the Bacillus-Subtilis Acetate Kinase Gene by Ccpa. *Journal of bacteriology, 175*(22), 7348-7355.

Halstead, J. R., Vercoe, P. E., Gilbert, H. J., Davidson, K., & Hazlewood, G. P. (1999). A family 26 mannanase produced by *Clostridium thermocellum* as a component of the cellulosome contains a domain which is conserved in mannanases from anaerobic fungi. *Microbiol, 145*, 3101-3108.

Henkin, T. M. (1996). The role of the CcpA transcriptional regulator in carbon metabolism in *Bacillus subtilis*. *FEMS Microbiol Lett, 135*(1), 9-15. doi: 10.1111/j.1574-6968.1996.tb07959.x

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review, 106*(4), 620-630. doi: 10.1103/PhysRev.106.620

Kahel-Raifer, H., Jindou, S., Bahari, L., Nataf, Y., Shoham, Y., Bayer, E. A., . . . Lamed, R. (2010). The unique set of putative membrane-associated anti-Sigma factors in *Clostridium thermocellum* suggests a novel extracellular carbohydrate-sensing mechanism involved in gene regulation. *FEMS microbiology letters, 308*(1), 84-93. doi: 10.1111/j.1574-6968.2010.01997.x

Kurokawa, J., Hemjinda, E., Arai, T., Kimura, T., Sakka, K., & Ohmiya, K. (2002). *Clostridium thermocellum* cellulase CelT, a family 9 endoglucanase without an Ig-like domain or family 3c carbohydrate-binding module. *Applied Microbiology and Biotechnology, 59*(4-5), 455-461. doi: 10.1007/s00253-002-1048-y

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., . . . Higgins, D. G. (2007). Clustal W and clustal X version 2.0. *Bioinformatics, 23*(21), 2947-2948. doi: 10.1093/bioinformatics/btm404

Lesburg, C. A., Cable, M. B., Ferrari, E., Hong, Z., Mannarino, A. F., & Weber, P. C. (1999). Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nature Structural Biology, 6*(10), 937-943.

Linville, J. L., Rodriguez, M., Land, M., Syed, M. H., Engle, N. L., Tschaplinski, T. J., . . . Cox, C. D. (2013). Industrial Robustness: Understanding the Mechanism of Tolerance for the Populus Hydrolysate-Tolerant Mutant Strain of Clostridium thermocellum. *Plos One, 8*(10). doi: UNSP e78829
DOI 10.1371/journal.pone.0078829

Lynd, L. R., Weimer, P. J., van Zyl, W. H., & Pretorius, I. S. (2002a). Microbial Cellulose Utilization: Fundamentals and Biotechnology. *Microbiol Mol Biol Rev, 66*(3), 506-577. doi: 10.1128/mmbr.66.3.506-577.2002

Lynd, L. R., Weimer, P. J., van Zyl, W. H., & Pretorius, I. S. (2002b). Microbial Cellulose Utilization: Fundamentals and Biotechnology. *Microbiology and Molecular Biology Reviews, 66*(3), 506-577. doi: 10.1128/mmbr.66.3.506-577.2002

Lynd, L. R., Wyman, C. E., & Gerngross, T. U. (1999). Biocommodity Engineering. *Biotechnology Progress, 15*(5), 777-793. doi: 10.1021/bp990109e

Lynd, L. R., Zyl, W. H. v., McBride, J. E., & Laser, M. (2005). Consolidated bioprocessing of cellulosic biomass: an update. *Current Opinion in Biotechnology, 16*(5), 577-583. doi: http://dx.doi.org/10.1016/j.copbio.2005.08.009

Maki, M., Leung, K. T., & Qin, W. S. (2009). The prospects of cellulase-producing bacteria for the bioconversion of lignocellulosic biomass. *IntJ Bioll Sci, 5*(5), 500-516.

Mao, X. Z., Ma, Q., Zhou, C., Chen, X., Zhang, H. Y., Yang, J. C., . . . Xu, Y. (2014). DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Research, 42*(D1), D654-D659. doi: Doi 10.1093/Nar/Gkt1048

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376-380. doi: Doi 10.1038/Nature03959

Mussatto, S. I., Dragone, G., Guimarães, P. M. R., Silva, J. P. A., Carneiro, L. M., Roberto, I. C., . . . Teixeira, J. A. (2010). Technological trends, global market, and challenges of bio-ethanol production. *Biotechnology Advances, 28*(6), 817-830. doi: http://dx.doi.org/10.1016/j.biotechadv.2010.07.001

Nataf, Y., Bahari, L., Kahel-Raifer, H., Borovok, I., Lamed, R., Bayer, E. A., . . . Shoham, Y. (2010a). *Clostridium thermocellum* cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors. *Proceedings of the National Academy of Sciences of the United*

States of America, 107(43), 18646-18651. doi: Doi 10.1073/Pnas.1012175107
Nataf, Y., Bahari, L., Kahel-Raifer, H., Borovok, I., Lamed, R., Bayer, E. A., . . . Shoham, Y. (2010b). *Clostridium thermocellum* cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors. *Proc Natl Acad Sci USA, 107*(43), 18646-18651. doi: Doi 10.1073/Pnas.1012175107
Nataf, Y., Yaron, S., Stahl, F., Lamed, R., Bayer, E. A., Scheper, T. H., . . . Shoham, Y. (2009a). Cellodextrin and Laminaribiose ABC Transporters in Clostridium thermocellum. *Journal of Bacteriology, 191*(1), 203-209. doi: Doi 10.1128/Jb.01190-08
Nataf, Y., Yaron, S., Stahl, F., Lamed, R., Bayer, E. A., Scheper, T. H., . . . Shoham, Y. (2009b). Cellodextrin and Laminaribiose ABC Transporters in *Clostridium thermocellum*. *J Bacteriol, 191*(1), 203-209. doi: Doi 10.1128/Jb.01190-08
Newcomb, M., Chen, C.-Y., & Wu, J. H. D. (2007a). Induction of the *celC* operon of *Clostridium thermocellum* by laminaribiose. *Proc. Natl Acad Sci USA, 104*(10), 3747-3752. doi: 10.1073/pnas.0700087104
Newcomb, M., Chen, C.-Y., & Wu, J. H. D. (2007b). Induction of the *celC* operon of *Clostridium thermocellum* by laminaribiose. *Proceedings of the National Academy of Sciences of the United States of America, 104*(10), 3747-3752. doi: 10.1073/pnas.0700087104
Newcomb, M., Millen, J., Chen, C.-Y., & Wu, J. H. D. (2011a). Co-transcription of the *celC* gene cluster in *Clostridium thermocellum*. *Appl Microbiol Biotechnol, 90*(2), 625-634. doi: 10.1007/s00253-011-3121-x
Newcomb, M., Millen, J., Chen, C.-Y., & Wu, J. H. D. (2011b). Co-transcription of the *celC* gene cluster in *Clostridium thermocellum*. *Applied Microbiology and Biotechnology, 90*(2), 625-634. doi: 10.1007/s00253-011-3121-x
Olson, D. G., & Lynd, L. R. (2012). Chapter seventeen - Transformation of *Clostridium thermocellum* by Electroporation. In J. G. Harry (Ed.), *Methods in Enzymology* (Vol. Volume 510, pp. 317-330): Academic Press.
Puri-Taneja, A., Paul, S., Chen, Y. H., & Hulett, F. M. (2006). CcpA causes repression of the phoPR promoter through a novel transcription start site, P-A6. *Journal of bacteriology, 188*(4), 1266-1278. doi: Doi 10.1128/Jb.188.4.1266-1278.2006
Raman, B., McKeown, C. K., Rodriguez, M., Jr., Brown, S. D., & Mielenz, J. R. (2011). Transcriptomic analysis of *Clostridium thermocellum* ATCC 27405 cellulose fermentation. *BMC Microbiol, 11*, 134. doi: 10.1186/1471-2180-11-134
Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., & Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research, 32*, D91-D94. doi: Doi 10.1093/Nar/Gkh012

Sannigrahi, P., Ragauskas, A. J., & Tuskan, G. A. (2010). Poplar as a feedstock for biofuels: A review of compositional characteristics. *Biofuels, Bioproducts and Biorefining, 4*(2), 209-226. doi: 10.1002/bbb.206
Schmittgen, T. D., & Livak, K. J. (2008). Analyzing real-time PCR data by the comparative CT method. *Nat Protocols, 3*(6), 1101-1108.
Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods, 9*(7), 671-675. doi: Doi 10.1038/Nmeth.2089
Schneider, T. D. (1997). Information content of individual genetic sequences. *Journal of Theoretical Biology, 189*(4), 427-441. doi: 10.1006/jtbi.1997.0540
Schneider, T. D., & Stephens, R. M. (1990). Sequence Logos - a New Way to Display Consensus Sequences. *Nucleic Acids Research, 18*(20), 6097-6100. doi: Doi 10.1093/Nar/18.20.6097
Schneider, T. D., Stormo, G. D., Gold, L., & Ehrenfeucht, A. (1986). Information-Content of Binding-Sites on Nucleotide-Sequences. *Journal of Molecular Biology, 188*(3), 415-431.
Schumacher, M. A., Seidel, G., Hillen, W., & Brennan, R. G. (2006). Phosphoprotein Crh-Ser(46)-P displays altered binding to CcpA to effect carbon catabolite regulation. *Journal of Biological Chemistry, 281*(10), 6793-6800. doi: Doi 10.1074/Jbc.M509977200
Schwarz, W. H. (2001). The cellulosome and cellulose degradation by anaerobic bacteria. *Appl Microbiol Biotechnol, 56*(5-6), 634-649.
Sheehan, J., & Himmel, M. (1999). Enzymes, energy, and the environment: A strategic perspective on the US Department of Energy's Research and Development Activities for Bioethanol. *Biotechnology Progress, 15*(5), 817-827. doi: 10.1021/bp990110d
Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics, 31*(1), 64-68. doi: Doi 10.1038/Ng881
Sierro, N., Makita, Y., de Hoon, M., & Nakai, K. (2008). DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Research, 36*, D93-D96. doi: Doi 10.1093/Nar/Gkm910
Solomon, B. D., Barnes, J. R., & Halvorsen, K. E. (2007). Grain and cellulosic ethanol: History, economics, and energy policy. *Biomass & Bioenergy, 31*(6), 416-425. doi: Doi 10.1016/J.Biombioe.2007.01.023
Stormo, G. D. (1998). Information Content and Free Energy in DNA–Protein Interactions. *Journal of Theoretical Biology, 195*(1), 135-137. doi: http://dx.doi.org/10.1006/jtbi.1998.0785
Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. *Nature, 403*(6765), 41-45.
Tripathi, S. A., Olson, D. G., Argyros, D. A., Miller, B. B., Barrett, T. F., Murphy, D. M., . . . Caiazza, N. C. (2010). Development of pyrF-Based Genetic

System for Targeted Gene Deletion in *Clostridium thermocellum* and Creation of a *pta* Mutant. *Appl Environ Microbiol, 76*(19), 6591-6599. doi: 10.1128/aem.01484-10

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research, 40*(15). doi: ARTN e115

DOI 10.1093/nar/gks596

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., . . . Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods, 5*(9), 829-834. doi: Doi 10.1038/Nmeth.1246

Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S., & Cantley, L. C. (2001). A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nature Biotechnology, 19*(4), 348-353. doi: Doi 10.1038/86737

Yang, S., Giannone, R., Dice, L., Yang, Z., Engle, N., Tschaplinski, T., . . . Brown, S. (2012). *Clostridium thermocellum* ATCC27405 transcriptomic, metabolomic and proteomic profiles after ethanol stress. *BMC Genomics, 13*(1), 336.

Zhang, Y. H. P., & Lynd, L. R. (2005). Regulation of cellulase synthesis in batch and continuous cultures of *Clostridium thermocellum*. *J Bacteriol, 187*(1), 99-106. doi: Doi 10.1128/Jb.187.1.99-106.2005

VITA

Jinlyung Choi was born in Incheon, South Korea to the parents of Keun-sik Choi and Soon-ok Seo. He has an order sister, Joo-hyung Choi. He attended Hak-ik High school. He also attended Inha University at Incheon South Korea majoring in Biotechnology Engineering. He also has minor in business through Techno-MBA program. In his freshman year he joined Incom computer club. He obtained his Bachelor's of Science degree from the Inha University in February 2008 in Biotechnology Engineering. He accepted a graduate research assistantship at the University of Tennessee, Knoxville, in the Department of Chemical and Biomolecular Engineering. He is married to Lanying Ma in 2012. His daughter Arielle was born in March 2014. Jinlyung Choi graduated with a Doctorate of Philosophy in Chemical and Biomolecular Engineering in May 2015.